NON-DESTRUCTIVE ASSESSMENT OF CHICKEN EGG FERTILITY USING HYPERSPECTRAL IMAGING TECHNIQUE

By

Adeyemi Olutoyin Adegbenjo

Department of Bioresource Engineering
McGill University, Montreal
Canada

December 2018

A Thesis Submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

DEDICATION

This thesis is dedicated to the Lord God Almighty, who has ever been the source of my strength and inspiration. May I never forget Him all the days of my life's sojourn. Indeed, I would have been consumed during my PhD journey but for His mercies. Also, to my Aunt (Mrs E.A. Eweoya), who just passed on while writing this dedication, and likewise to my senior friend and fellow soldier of the cross (Dr. S.O. Olanisebe), who also dropped the baton of life shortly after my Aunt. Lastly, to a brave fighter and loving brother (Emeka Ngadi), who passed on at the point of submitting this thesis! I owe this generation a promise as the Lord enables me, which is to commit to the course and researches into early detection and diagnosis via machine learning and artificial intelligence.

ABSTRACT

The Canadian chicken industry is a huge one with about 2,836 regulated producers spread across the provinces producing, and of which 61% of production originated from Quebec and Ontario. According to the Agriculture and Agri-food Canada report 2017, total hatching egg set (for both egg production chicks and broilers) was over 1.0 billion. With fertility rate observed in the year 2017 to be around 82%, there were about 180 million unhatched eggs incubated in Canada for year 2017 alone. This meant a whooping sum of at least 311 million Canadian dollars was wasted by the hatchery industries towards incubating unhatched eggs for the year 2017. Whereas, this non-hatching, non-fertile eggs can find useful applications as commercial table eggs or low-grade food stock if they can be detected early and isolated accordingly, especially prior to incubation. The primary goal of this research is to investigate the use of a near infrared (NIR) hyperspectral imaging (HSI) technique in a non-destructive assessment of early chicken egg fertility recognition and discrimination.

The first study examined the suitability of a chemometric partial least square (PLS) regression algorithm, towards building a robust model for objective prediction of chicken egg fertility. A moving-threshold technique was implemented for discrimination based on PLS regression results on the calibration set. For the brown eggs on considered incubation days 0 to 4, true positive rates (TPR) ranged from 95.65% to 100% and true negative rates (TNR) ranged from 88.10% to 93.57%. White eggs on the other hand has true positive rates (TPR) ranging from 95.24% to 100% and true negative rates (TNR) ranging from 91.35% to 95.83%. All results were obtained at selected threshold values of between 0.50-0.85. The results indicated that the adapted PLS regression technique can accurately discriminate between fertile and non-fertile eggs, prior to incubation and on different days of incubation. It was further established that despite the PLS regression approach worked for the chicken egg classification task, the results were promising with the use of many

PLS components (PCs), but the use of fewer PCs shifted classification accuracies in favour of the prevalent class due to the imbalance data structure phenomenon. It therefore became imperative to improve on the present implementation mode of PLS for classification algorithm, in a view to accommodating the chicken egg fertility data structure and at the same time allowing the use of adequate number of PCs. Based on the present results, the second study tested the appropriateness of a PLSDA learner-based feature selection algorithm, with a non-parametric receiver operating characteristic (ROC) curve analysis technique for identifying informative features, towards improving model performance for early chicken egg fertility classification. Data were first resampled following a matched nested case-control study approach. With only a maximum number of 5 PCs considered, classifier performance greatly improved with selected ratio features; having TPR, TNR, and AUC (area under ROC curve) values in the range of 90-100%, obtained prior to and on different days of incubation. Chicken egg fertility model structure was eventually successfully developed, validated, and verified using maximum optimum number of 3 PCs.

Understanding that the modelling approach used to identify informative variables might not be the best approach to translate the identified features into Industrial practice, 10 different classifier performances were compared and contrasted in the third study for adoptability potentials towards building an industrial online chicken egg fertility assessment system. From the sensitivity, specificity, precision, and F1-score values of 100.00%, 87.00%, 93.80%, and 96.80% respectively for brown eggs and 100.00%, 71.40%, 87.80%, and 93.50% respectively for white eggs, the k-nearest neighbours (KNN) classifier was adjudged preferable above its other counterparts including partial least squares discriminant analysis (PLSDA), sparse PLSDA (sPLSDA), orthogonal PLSDA (OPLSDA), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Mahalanobis discriminant analysis (MDA), support vector machine (SVM),

random forest (RF), logistic regression (LOGREG), and soft independent modelling of class analogy (SIMCA). The final study seeks to verify the reproducibility of resampling preprocessing methodology by examining the performance of an automatic synthetic minority oversampling technique (SMOTE) algorithm on a fairly large industrial scale (10, 000) chicken egg fertility data set. KNN classifier already presented as optimal among other classifiers was used for discrimination and performance evaluated from sensitivity (SEN- 92.10%), specificity (SPE- 80.50%), precision (PPV- 99.10%), area under ROC curve (AUC- 91.70%), and overall accuracy (OVA- 91.60%). Our latest results based on the considered evaluation criteria were comparable with previous results, showing reproducibility potential of our methodology.

RÉSUMÉ

L’industrie canadienne de volailles est très importante avec environ 2836 producteurs enregistrés étendus à travers les provinces canadiennes, en fait, 61% de la production provient du Québec et de l’Ontario. Si on se base sur le rapport fait par Agriculture et Agroalimentaire Canada en 2017, le total d’œufs à couver (Pour la production d’œufs à pondeuses ou des poulets de chair) était de plus de 1,0 milliard. En plus, le niveau de fertilité observé pour 2017 était autour de 82%, Il y avait près de 180 millions d’œufs non éclos au Canada seulement en 2017. Cela signifiait que l’industrie de l'écloserie avait gaspillé au moins 311 millions de dollars canadiens en incubation d'œufs non éclos pour l'année 2017. Alors que ces œufs non couvés et non fertiles peuvent trouver des applications utiles en tant qu’œufs de consommation commerciale ou réserves d’aliments de qualité médiocre s’ils peuvent être détectés tôt et isolés en conséquence, en particulier avant l’incubation. L'objectif principal de cette recherche est d'étudier la pertinence d'une technique d'imagerie hyperspectrale (HSI) dans la région proche infrarouge (NIR), une méthode de contrôle et de reconnaissance non destructive à la fertilité des œufs de poule précoce.

La première étude a examiné la pertinence d'une analyse chimiométrique partielle par régression des moindres carrés (PLS) en vue de la création d'un modèle robuste de prédiction objective de la fertilité des œufs de poule. La technique de seuil mobile a été mise en œuvre pour la discrimination basée sur les résultats de la régression PLS sur le jeu d'étalonnage. Avec des taux positifs réels (TPR) allant jusqu'à 100% et une correspondance acceptable des taux vrais négatifs (TNR), obtenus à des valeurs de seuil sélectionnées comprises entre 0,50 et 0,85; les résultats ont indiqué que la technique de régression PLS adaptée peut discriminer avec précision les œufs fertiles et non fertiles, avant l'incubation et à différents jours d'incubation. Il a été établi que, malgré l’approche de régression PLS adaptative appliquée à la tâche de classification des œufs de poule,

les résultats ont été excellents avec l'utilisation de plusieurs composants PLS (PCs), l'utilisation de moins de PCs a modifié les précisions de classification en faveur de la classe dominante en raison du phénomène de structure de données de déséquilibre. Il devenait donc impératif d'améliorer le mode de mise en œuvre actuel de PLS pour l'algorithme de classification, afin de tenir compte de la structure de données sur la fertilité des œufs de poule tout en permettant l'utilisation d'un nombre adéquat de PCs. Sur la base des résultats actuels, la seconde étude a testé la pertinence d'un algorithme PLS-DA de sélection de caractéristiques basé sur l'apprenant, avec une technique d'analyse de courbe non-paramétrique caracteristique de fonctionnement de récepteur *(ROC)* permettant d'identifier des caractéristiques informatives, dans le but d'améliorer les performances du modèle pour la classification de la fertilité précoce des œufs de poule. Les données ont d'abord été rééchantillonnées selon une approche d'étude cas-témoins appariée. Avec un nombre maximal de 5 PCs considérer, les performances du classificateur ont été considérablement améliorées avec les fonctionnalités de caractéristiques sélectionnées; ayant des valeurs de TPR, TNR et AUC (aire sous courbe ROC) comprises entre 90 et 100%, obtenues avant et pendant différents jours d'incubation. La structure du modèle de fertilité des œufs de poule a finalement été développée, validée et vérifiée avec succès en utilisant un nombre adéquat de PC.

Comprenant que l'approche de modélisation utilisée pour identifier les variables informatives n'était peut-être pas la meilleure solution pour traduire les caractéristiques identifiées en pratiques industrielles, 10 performances de classificateur différentes ont été comparées dans la troisième étude sur les potentiels d'adaptabilité vers la création d'un système d'évaluation industriel en ligne de la fertilité des œufs de poule en ligne. D'après les calculs de sensibilité, de spécificité, de précision, de score F1 et de ROC, le classificateur *KNN* a été jugé préférable à ses homologues, incluant le *PLSDA, sPLSDA, OPLS-DA, LDA, QDA, MDA, SVM, RF, LOG_REG,* et le *SIMCA.*

La dernière étude tend à vérifier la reproductibilité de la méthodologie de pré-traitement de rééchantillonnage en examinant les performances d'un algorithme automatique d'échantillonnage synthétique minoritaire *(SMOTE)* sur un ensemble de données sur la fertilité des œufs de poule à une échelle industrielle (10 000). Le classificateur KNN déjà présenté comme optimal parmi d'autres classificateurs a été utilisé pour la discrimination et les performances évaluées à partir de la sensibilité (SEN- 92,10%), de la spécificité (SPE- 80,50%), de la précision (PPV- 99,10%), de l'aire sous la courbe ROC (AUC- 91,70%), et précision globale (OVA- 91,60%). Nos derniers résultats fondés sur les critères d'évaluation retenus étaient comparables aux résultats précédents, montrant le potentiel de reproductibilité de notre méthodologie.

ACKNOWLEDGEMENTS

My appreciation goes to all that contributed to the success of this academic pursuit, for their immense assistance and useful contributions. I specially appreciate my supervisor, Prof. Michael Ngadi for accepting me into his hyperspectral imaging lab to undertake this study. Your unrelenting effort in seeing that this work is concluded and your guidance at critical stages of the work are well appreciated.

I thankfully acknowledge my Ph.D. committee members; Dr. Ashraf Ismail (Food Science and Agricultural Chemistry, McGill), Dr. Shiv Prasher (Bioresource Engineering, McGill), and Dr. Raj Duggavathi (Animal Science, McGill), for their suggestions and constructive criticisms for the improvement of my research proposal.

I am also grateful to all staff members of the Bioresouce Engineering Department, including but not limited to Dr. Valérie Orsat, Dr. Vijaya Raghavan, and Dr. Marie-Josée Dumont, for all support offered during my PhD programme at McGill. Special thanks to Ms Patricia Singleton, Mrs. Abida Subhan, Ms. Susan Gregus and Ms. Christiane Trudeau for all the administrative supports. I acknowledge the technical support provided by Dr. Samson Sotocinal, Mr. Scott Manktelow and Yvan Gariepy at some experimental design stages of the work.

My unreserved gratitude goes to my parents, Mr. and Mrs. A. A. Adegbenjo. To you I owe my basic foundational knowledge in Mathematics and English Language. It is my prayer that you will live long seeing the outcome of the travails of your souls and being satisfied. Likewise, my siblings: Dr. Adewale Adegbenjo, Dr. Adedotun Adegbenjo, Adedayo Adegbenjo and Adedokun Adegbenjo, thanks for your emotional support and understanding throughout the course of this study.

To my friends and research group members: Dr. Ebenezer Kwofie, Dr. Peter Adewale, Dr. Olanike Aladenola, Dr. Alaba Boluwade, Dr. Ogan Mba, Dr. Feifei Tao, Dr. Hui Huang, Dr. Laura Liu, Dr. Nandkishor Dhawale, Dr. Senthilkumar Thiruppathi, Hernán Rey Sánchez Solano, Patrick Cortbaoui, Chijioke Ejebe, Emmanuella Ellis, Audrey Yank, Chijioke Nwankpa, Wathsala Tennakoon, Mitalie Makhani, Tebogo Leepile, Nnedimma Nnebe, Christopher Kucha, Christopher Nzediegwu, Tosin Oludare, Jacob Liberty, Babatunde Onadipe, Lanre Adetunji, Omotola Folarin-Ottun, Olanike Oyeleke, Ademola Adekunle and Omotola Oyekunle. You are all acknowledged for your encouragement, helpful comments, and supports.

I must but appreciate my spiritual family including the youth group and the entire members of the Redemption Bible Church, Montreal QC, CA, for empowering me during the study period. The spiritual backbone and strength received from partnering with the discipleship labours of the Peace House Ministries, Gboko, Nigeria cannot be overemphasised.

My darling, Olubusola was always lifting my hands up in prayers and encouragement when at many cross roads along the line. You will ever be my choice, love. I appreciate our children: Godlymodel, Godlycharacter, and Godlyvirtue for coping with a student dad. I cannot love you enough.

I am grateful to the following organisations for financial supports: McGill University, Canada; Obafemi Awolowo University, Nigeria, TETFUND, Nigeria.

Above all, I am infinitely grateful to my Lord Jesus Christ. Without Him, I can do nothing. He gives power to the weak, strength to the feeble and inspiration to them that lacks it. You gave me the wisdom to cope throughout this study period, even when I could have quitted. You have made me whom I am today. To you alone be all the glory and adoration.

CONTRIBUTION OF AUTHORS

Adeyemi Olutoyin Adegbenjo is the main author of this work, supervised by Dr. Michael Ngadi from the Department of Bioresource Engineering, McGill University, Sainte Anne-de-Bellevue, Quebec, Canada.

Dr. Michael Ngadi, the supervisor and major advisor on the thesis, co-authored all manuscripts, and provided scientific guidance in the planning and execution of the work as well as co-editing and reviewing manuscripts.

Dr. Li Liu co-authored the third and sixth chapters and provided technical assistance during analysis and testing of models. She also made contribution in reviewing some of the manuscripts.

Details of the manuscripts to be submitted for publication are as shown below:

A. Journal papers

1. **Adegbenjo, A. O.,** and Ngadi, M. (2019). Towards Improvement in Non-destructive Assessment of Early Chicken Egg Fertility Discrimination: A Review. *Sensors* (based on literature review Chapter 2 of the thesis).
2. **Adegbenjo, A. O.,** Liu, L., and Ngadi, M. (2019). An adaptive partial least square (PLS) regression approach for classifying chicken egg fertility hyperspectral imaging data. *Chemometrics and Intelligent Laboratory Systems* (Chapter 3 of the thesis).
3. **Adegbenjo, A. O.,** and Ngadi, M. (2019). A non-parametric ratio-based feature selection approach for chicken egg fertility classification using hyperspectral imaging. *IEE transactions on Pattern Analysis and Machine Intelligence* (Chapter 4 of the thesis).

4. **Adegbenjo, A. O.,** and Ngadi, M. (2019). Visualization of Hyperspectral Imaging (HSI) Based Chicken Egg Fertility Data for Early Discrimination Using Combination of Classifiers. *Information Sciences* (Chapter 5 of the thesis).
5. **Adegbenjo, A. O.,** Liu, L., and Ngadi, M. (2019). Improved Chicken Egg Fertility Classification Using SMOTE Preprocessing Algorithm and K-Nearest Neighbours' Classifier. *Expert Systems with Applications* (Chapter 6 of the thesis).

B. Papers presented at scientific and technical conferences

1. **Adegbenjo, A. O.,** and Ngadi M., (2017). Preprocessing techniques for hyperspectral imaging data. Oral presentation at the *CSBE/SCGAB Technical Conference & AGM,* from August 6 - 10, 2017, Winnipeg, Manitoba, Canada.
2. **Adegbenjo, A. O.,** Liu, L., and Ngadi M., (2017). Spectral characteristics of chicken egg shell and yolk. Oral presentation at the *Northeast Agricultural & Biological Engineering Conference (NABEC) Meeting, July 30-August 2, 2017,* Kemptville, Groton, CT, USA.
3. **Adegbenjo, A. O.,** Liu, L., and Ngadi M., (2016). An adaptive PLS approach for discriminating egg fertility data using hyperspectral images. Oral presentation at the *International Union of Food Science and Technology (IUFoST) Conference, from August 21 - 25, 2016,* Dublin Ireland.
4. **Adegbenjo, A. O.,** and Ngadi M., (2015). Handling imbalanced data problems in Food processing analysis. Postal presentation at the *CSBE/SCGAB Technical Conference & AGM,* from July 5 - 8, 2015, Edmonton, AB, Canada.
5. **Adegbenjo, A. O.,** Liu, L., and Ngadi M., (2015). A Non-destructive evaluation of eggshell strength during incubation, using hyperspectral imaging. Oral presentation at

Oral presentation at the *12th International Congress on Engineering and food, (ICEF 12)* from June 14 - 18, 2015, Quebec City, QC, Canada.

6. **Adegbenjo, A. O.,** Ngadi, M., and Liu, L., (2014). Effect of eggshell pigmentation on hyperspectral properties and characteristics of brown chicken eggs. Oral presentation at the 2014 *Northeast Agricultural & Biological Engineering Conference (NABEC) Meeting,* July 27-30, Kemptville, Ontario, Canada.
7. **Adegbenjo, A. O.,** Ngadi, M., and Liu, L., (2014). Near Infrared Hyperspectral Imaging for Early Prediction of Fertility in Chicken Eggs. Oral presentation at the 2014 *Annual International Meeting of the ASABE*, July 13-16, Montreal, QC, Canada.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

AUC	Area Under ROC Curve
CV	Cross Validation
FESEM	Field Emission Scanning Electron Microscope
FOCSVM	Fuzzy One Class Support Vector Machine
ICA	Independent Component Analysis
KNN	K Nearest Neighbours'
LDA	Linear Discriminant Analysis
LOGREG	Logistic Regression
MCCV	Monte Carlo Cross Validation
MDA	Mahalanobis Discriminant Analysis
MRI	Magnetic Resonance Imaging
NPV	Negative Predictive Value
OPLSDA	Orthogonal Partial Least Squares Discriminant Analysis
OVA	Overall Accuracy
PCA	Principal Component Analysis
PCs	Principal Components/PLS components
PLS	Partial Least Squares
PLSDA	Partial Least Squares Discriminant Analysis
PPV	Positive Predictive Value
PRE	Precision
QDA	Quadratic Discriminant Analysis
RF	Random Forest
ROC	Receivers Operating Characteristic Curve
SEN	Sensitivity
SIMCA	Soft Independent Modelling of Class Analogy
SMOTE	Synthetic Minority Oversampling Technique
SPE	Specificity
sPLSDA	Sparse Partial Least Squares Discriminant Analysis
SVM	Support Vector Machine

TNR	True Negative Rate
TPR	True Positive Rate
TR	Threshold

CHAPTER 1

INTRODUCTION

1.1 Background

The Canadian chicken industry is a huge one with about 2,836 regulated producers spread across the provinces producing. There are about 241 broiler hatching egg producers and about 1,059 egg producers. Canada produced up to 1.2 billion kilograms of chicken in 2017, of which 61% of production originated from Quebec and Ontario (Agriculture and Agri-Food Canada, 2017). In 2017, Canada exported about 14.7 million chicks and poults (worth $56.0 million), 39.8 million hatching eggs, worth 68.8 million dollars as against 22.7 million hatching eggs in 2013 (worth $36 million). According to the Agriculture and Agri-food Canada report 2017, total hatching egg set (for both egg production chicks and broilers) was over 1.0 billion. With fertility rate observed in the year 2017 to be around 82%, there were about 180 million unhatched eggs incubated in Canada for year 2017 alone. This meant a whooping sum of at least 311 million Canadian dollars was wasted by the hatchery industries towards incubating unhatched eggs for the year 2017. Whereas, this non-hatching, non-fertile eggs can find useful applications as commercial table eggs or low-grade food stock if they can be detected early and isolated accordingly, especially prior to incubation.

Discarding of non-hatching eggs has consistently posed significant disposal problems for the hatcheries, especially in the case of exploder eggs in hatching cabinet, resulting in high tendency of molds and bacteria infestation to other eggs (Lawrence et al., 2006). Thus, identification and isolation of infertile eggs have significant economic and safety implications for commercial broiler breeders.

Conventional method of chicken egg fertility assessment termed candling is subjective, cumbersome, and slow. Also, because eggs are selected randomly for candling, majority of eggs escapes being tested, making the candling system to be eventually inefficient. Apart from the candling system being laborious and inaccurate, it is also not appropriate for building an online egg fertility classification system, in a fast pace technology advancing era of our days. Hence, there is a need for a non-destructive, fast and online prediction technology to assist with early chicken egg fertility identification problem. Electromagnetic radiation-based approaches have therefore been identified as suitable for filling the identified gaps with chicken egg fertility assessment.

When electromagnetic radiation hits an object, the output of this interaction is dependent on the properties of the object such as colour, physical damage and presence of foreign material on the object`s surface. Various types of electromagnetic radiation have been known to be used in quality control of foods. For example, near infrared radiation have been widely used to measure moisture content and internal defects have been known to be detected using X-rays (Sahin & Sumnu, 2006). The electromagnetic radiation is transmitted in the form of waves and it can be classified according to wavelength and frequency.

Spectroscopic methods provide detailed fingerprints of the biological sample to be analysed using physical characteristics of the interaction between electromagnetic radiation and the sample material, such as absorbance, transmittance, reflectance, fluorescence, phosphorescence, and radioactive decay (Mehl et al., 2004). The analytic spectral regions include the ultraviolet, visible, near-, mid-, and far- infrared regions. The near-infrared (NIR) regions has been successfully used for food quality and food safety analysis during the past two decades (Chen, 1993; Chen et al., 1998; Osborne et al., 1993; Williams & Norris, 1987).

Hyperspectral imaging (HSI) builds upon the merits and demerits of conventional imaging and spectroscopy techniques to produce a two-dimensional spatial array of vectors which represents the spectrum at each pixel position (ElMasry & Sun, 2010a). The final output of a hyperspectral image is a three-dimensional "hypercube" consisting of the two spatial dimensions and one spectral dimension. This is then analysed for products compositional identification and authentication (Chen et al., 2002; Kim et al., 2002; Mehl et al., 2004; Schweizer & Moura, 2001).

Recent researches indeed have supported the great potential applications of HSI as a non-destructive method for assessing hatchability, embryo development and mortality rates in chicken eggs. These studies have however reported mostly on fertility detection of white-shelled chicken eggs with scanty reports on brown eggs and where available, results were not as promising as with white eggs. Also, samples considered in earlier studies were small (Lawrence et al., 2006; Liu & Ngadi, 2013; Smith et al., 2008).

Smith et al. (2008) reported low validation and verification accuracies for fertility detection in brown eggs (Validation data sets: 71% for Day 0; 63% for Day 1, 65% for Day 2, 83% for Day 3; Verification data sets: 51% for Day 0 and 50% for Day 3). It was concluded that the Mahalanobis Distance (MD)/Principal Component Analysis (PCA) model used was not adequate for the discrimination. This is of a great concern for the poultry industry as this means large number of fertile eggs would end up being discarded based on such model. Therefore, there is indeed an urgent need for more appropriate discrimination technique(s) for egg fertility assessment.

Liu and Ngadi (2013) successfully used the HSI in the NIR wavelength region to detect fertility and embryo mortality in white leghorn eggs. With up to 100% accuracy obtained for fertility detection on Day 0, this finding strengthened the assertion that HSI has great potential applications in the poultry hatchery industry. Owing to the knowledge that improved assessment

of relevant chemical and other quality attributes are more readily obtained in the NIR region, the success of Liu and Ngadi (2013) have been linked to higher wavelength consideration and improved HSI technology. However, brown eggs were not considered in this study and validation of results with independent egg samples was also not done. In the same vein, non-supervised learning techniques of PCA and K-means clustering were used for classification. Unsupervised classification has been reported as a preliminary stage in any discrimination problem and should necessarily be followed by supervised classification (Swarbrick, 2012). Hence, non-supervised techniques usually need further confirmation using standard supervised learning approaches, towards a futuristic industrial adoptability.

Imbalanced data situation exists in most fields of endeavours like the biomedical, surveillance and security industries, insurance, management/finance (Artís et al., 2002; He & Garcia, 2009) and the Agricultural sectors. Specific cases in the Agricultural sector include fruit bruise detection, infectious fruit/vegetable prediction, and chicken egg fertility assessment. To the best of our knowledge, there have not been reported research efforts geared towards solving the imbalanced data situation in the Agricultural sector especially in food analysis research. Researchers in this field have been carrying out analysis and validating models on balanced data (Artís et al., 2002; Das & Evans, 1992b; Smith et al., 2008), the results of which when applied to the real-world situation are very prone to doubt (Kuhn & Johnson, 2016). This might indeed be the major reason for the low acceptability and adoptability of such models in real industrial settings. It is to this end that this research was set up to examine in the NIR region, early fertility assessment of brown and white chicken eggs, using suitable analytical techniques that will be more favourable in handling the imbalanced data problem which is the real situation with chicken egg fertility/early embryonic development detection.

1.2 Hypothesis

The theoretical background upon which this work was based is that hyperspectral information obtained from light spectra passing through an object can be linked directly with some important properties of the object in question. It is therefore hypothesized that the hyperspectral information obtained from chicken egg could be used for early prediction of egg fertility. It is also hypothesized that application of feature selection technique could improve accuracy of chicken egg fertility prediction. It is further hypothesized that the application of data preprocessing and/or recognition-based (one-class learning) approaches to imbalanced hyperspectral data could improve accuracy of results and adoptability of developed models. This research sought to strengthen existing knowledge of a great potential of hyperspectral imaging technique in an objective assessment of food products. It is believed that the techniques developed from this study will find industrial applications in the future real time and online detection/classification systems for food products.

1.3 Objectives

The general objective of this study is to investigate the use of NIR HSI technique in a non-destructive assessment of chicken egg fertility. One of the major tasks in this research is to come up with classification algorithm(s) suitable for the specific case of imbalanced data distribution with chicken egg fertility assessment. The outcome of this research will benefit tremendously the poultry industries, assisting greatly in maximizing profits when infertile eggs can be removed early, especially prior to incubation.

The Specific Objectives

The following specific objectives have been drawn towards accomplishing the stated general objective of this study:

1. To develop a model for objective prediction of fertility and /or early embryonic development of white and brown chicken eggs using hyperspectral imaging in conjunction with an adapted partial least square regression (PLSR) algorithm.
2. To examine the appropriateness of a PLSDA learner-based feature selection technique in improving model performance for early chicken egg fertility discrimination.
3. To compare the performances of some available classification algorithms in a way to determine which classifier or combination of classifiers might be well suited for translating the research outcome of early chicken egg fertility discrimination into industrial practice.
4. To evaluate the potential of optimizing predictive performance of chicken egg fertility classification models using SMOTE data preprocessing algorithm and KNN classifier

CHAPTER 2

LITERATURE REVIEW

2.1 Egg and egg production

Eggs have been a human food from the time immemorial, being one of nature's almost perfect source of protein and other high-valued nutrients. Eggs have found useful applications in both food industries and at home and of which chicken eggs are the most important. Eggs of other birds including but not limited to those of geese, ducks, plovers and quail are of lesser significance. Consequently, the term "egg" when used without prefix generally talks about chicken egg (Belitz et al., 2009). White eggs are produced from White Leghorn breed while brown eggs are produced from a hybrid of chicken including Rhode Island Red, Barred Plymouth Rock and New Hampshire.

Over 50 billion of chickens are being raised annually by poultry farmers all over the world be it as layers, towards egg production or as broilers, towards meat production, and production growth is anticipated to continue. The global world population has been projected to hit 9.6 billion by 2050, creating an increasing demand for animal-based food (Mottet & Tempio, 2017). Even though pork and beef demand could increase by up to 43% and 66% respectively, poultry meat has been projected to have greatest growth rate of up to 121%, and demand for eggs is expected to increase by 65% (Alexandratos & Bruinsma, 2012). For the year 2017, Canada exported over 39 million hatching eggs (worth over $68 million), with the US being the largest market. The importation figure for the same year however stood at over 141 million hatching eggs for broilers (worth over $49 million), with entire importation coming from the US (Agriculture and Agri-Food Canada report 2017).

Seeing the great importance of chicken and chicken eggs both locally and globally, it is imperative especially in the present era of advancing technologies in the field of machine learning and artificial intelligence, that there would be worthwhile assistance towards improving hatchability rate of chicken eggs. Early fertility and/or embryonic development detection would prevent wastage of egg and incubation energy, make more incubator space available for viable hatching eggs, and likewise promises huge economic returns. Achieving the above, would also put us on a good pedestal towards achieving the sustainable development goal 2 (SDG2) agendum.

2.1.1 Egg formation and structure

Egg formation is a process that occurs in about 25-26 hours from ovulation to oviposition. It commences with a matured ovum (which is a plain yolk and germinal disc) in the reproductive tract, resulting at last in a hard shelled egg, fully complete with its own protective membranes and the necessary nutrients needed for embryonic development (Latour et al., 2014). The major stages in egg formation include the ovulation stage, fertilization stage, formation and oviposition. All these stages are accomplished in the ovary and the oviduct. The process can be better and clearly understood considering the schematic diagram in Figure 2.1.

In the ovary, the ovum or oocyte is released from the follicle through a process known as ovulation. Ovulation takes place in about 5 to 10 minutes following the expulsion of the previous egg. This stage has been preceded with yolk production made possible from the chickens (hens) being fed with diets containing appropriate nutrients. Diets rich in calcium is of good necessity at this stage as it will find useful application later during the shell formation. These nutrients absorbed into the bloodstream from the hens' digestive tracts are converted into yolk by the hen's liver. The yolk is then transported through the blood stream from the liver to the ovary. Here, the follicular cells around the ovum take the yolk and other nutrients and carry them along to the ovum. The

immature ova and their neighboring follicular cells are securely embedded within the ovary. As the ovum increases in more and more yolk accumulation, it becomes greatly enlarged that it can no longer fit inside the ovary. Therefore, there begins a gradual continuous pushing of the nested ovum and the ovarian follicle towards the outer ovarian edge.

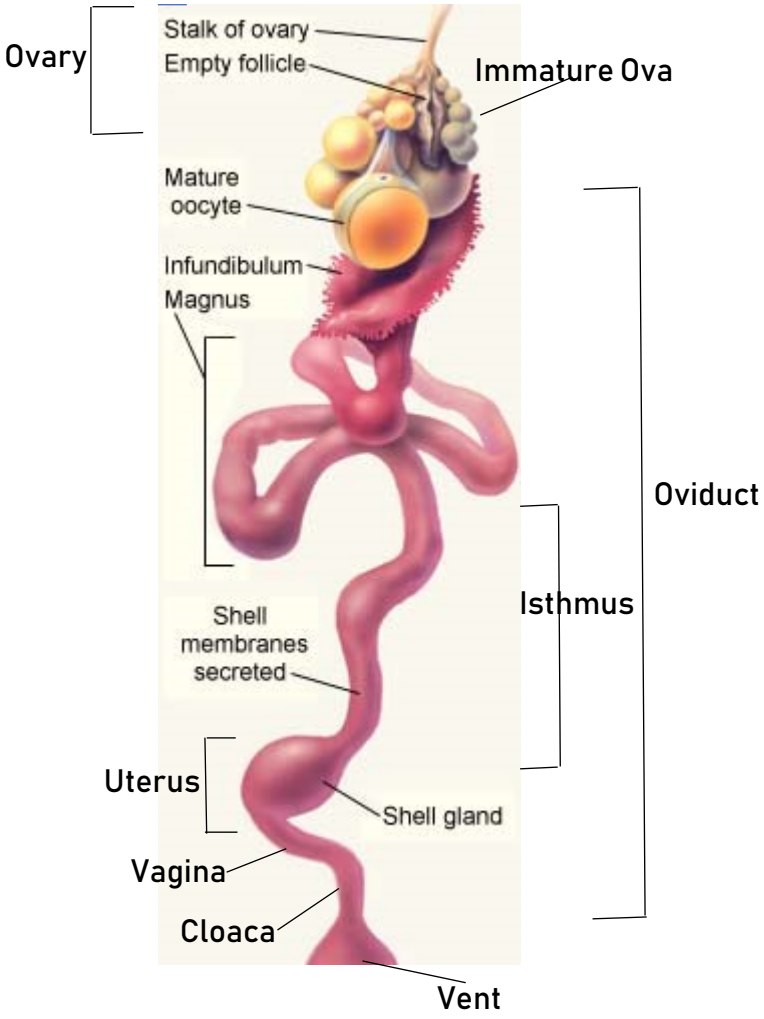


Figure 2.1. Schematic diagram detailing the process of egg formation

http://chickscope.beckman.uiuc.edu/resources/egg_to_chick/formation.html

As the ovum accumulates enough amount of yolk that is adequate for growing a chick, the ovum ruptures from its follicle through a process earlier mentioned as ovulation. The free ovum

drops into the ovarian pocket and within minutes is captured by the infundibulum and guided into the mouth of the hen's left oviduct. Just almost immediately the ovum is released from the ovary and before being received by the infundibulum, the egg's nucleus passes through a process of primary cell divisions known as meiosis and it is only one of the cells (others fade away naturally) produced from meiosis that ended up becoming a matured ovum that is accepted by the infundibulum. Fertilization occurs inside the infundibulum if sperm is available, and the resulting zygote thereby commences a secondary cell division via mitosis. The first layer of albumen is also deposited at this stage (Potter et al., 1998b). The remaining process of egg formation is completed during the journey down the oviduct.

The oviduct is divided into 6 different sections which are: infundibulum (oviduct's mouth or funnel), magnum, isthmus, shell gland (uterus), vagina and the cloaca. Whether the ovum is fertilized or not (as in the case of table or hatchery infertile eggs), it continues its journey along the oviduct to allow for complete covering by layers of egg white (albumen) and other internal supporting structures. The section of the oviduct responsible for most of albumen secretion is the magnum. The matured ovum and its surrounding layers reaching the magnum can now at this point be conveniently called an egg (if fertilised, an embryo is formed). Due to the spiral structural design of the oviduct, the egg twists/rotates along its journey and some protein fibers extension from the egg are hooked by the thick and thin albumens secreted along the oviduct. This occurrence results in albumen layers and chalazae formation. The shell membranes are then added in the Isthmus and the shell gland located in the uterus later commences the process of shell formation (Bruce et al., 1997). The average time as reported by (Coutts & Wilson, 2007; Sturkie, 1965) an ovum spends in each section as it travels down the oviduct are: infundibulum 15 minutes, magnum 2-3 hours, isthmus 1 hour, uterus 21 hours, vagina/cloaca just a few minutes. A finally formed

whole egg structure as shown in figure 2.2 consists of 30-33% yolk, about 60% albumen, and between 9-12% shell (Halls & Shur-Gain, 2014; Stadelman, 2017).

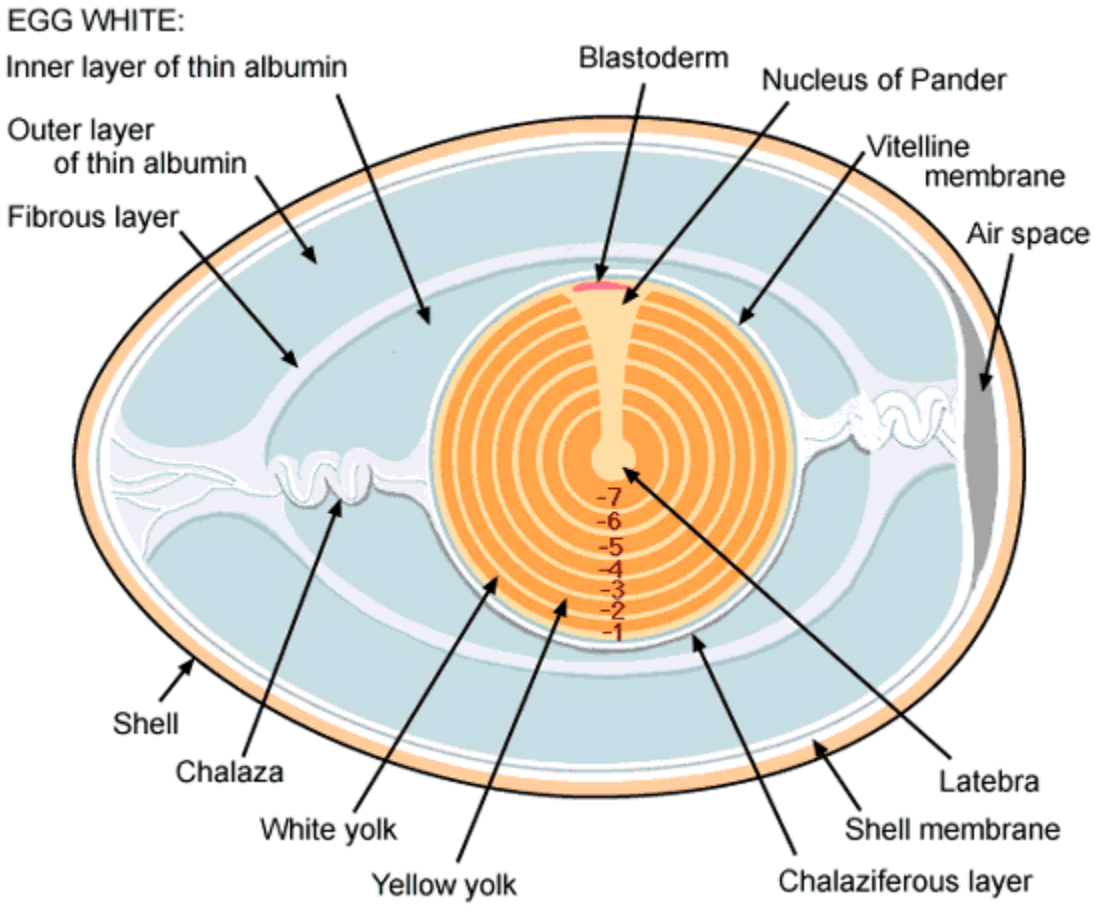


Figure 2.2. Chicken egg structure

http://chickscope.beckman.uiuc.edu/resources/egg_to_chick/formation.html

The eggshell is deposited while the egg is still in the hen's uterus. Three distinct stages of eggshell formation can be identified according to Hernández-Hernández et al. (2008) namely: (a) initial, (b) fast growth and (c) termination. The initial stage begins with calcium carbonate ($CaCO_3$) spheruliths forming on the eggshell membranes. This formation progresses until adjacent spheruliths are knitted (fused) together, a process known as nucleation. After this is an emergence of columnar crystals (palisades) from the spherules during the fast growth stage. Columnar crystal

formation continues until eggshell calcification is completed with the deposition of the cuticle layer in the termination stage. Figure 2.3 depicts the texture of the finally formed hen eggshell as imaged from a field emission scanning electron microscope (FESEM). It is important to mention that for brown eggs, deposit of protoporphyrin (pigment responsible for the brown colouration) occurs both at the onset and termination stage of shell formation.

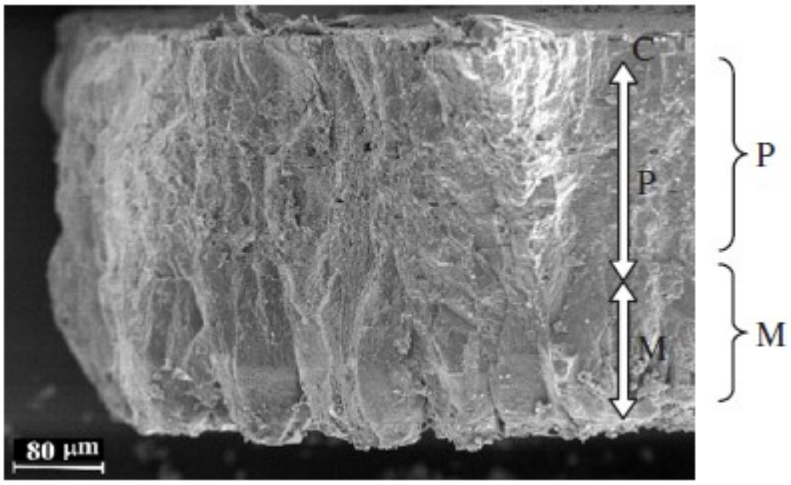


Figure 2.3 FESEM textural image of eggshell showing the mammillary layer (M), the palisade (P), and the cuticle (C). Source: (Hernández-Hernández et al., 2008)

After deposition of the cuticle (protective coating layer) over the shell in the termination stage of eggshell formation, oviposition (egg laying) is then initiated by hormonal contractions of the uterus and the egg is pushed through the vagina, down along the cloaca and exits through the vent. It takes about 24 hours for the egg to complete its journey through the oviduct. Eggs are usually laid during the middle of the day but if this is not completed until later in the day, the egg will remain at the end of the oviduct till the following day.

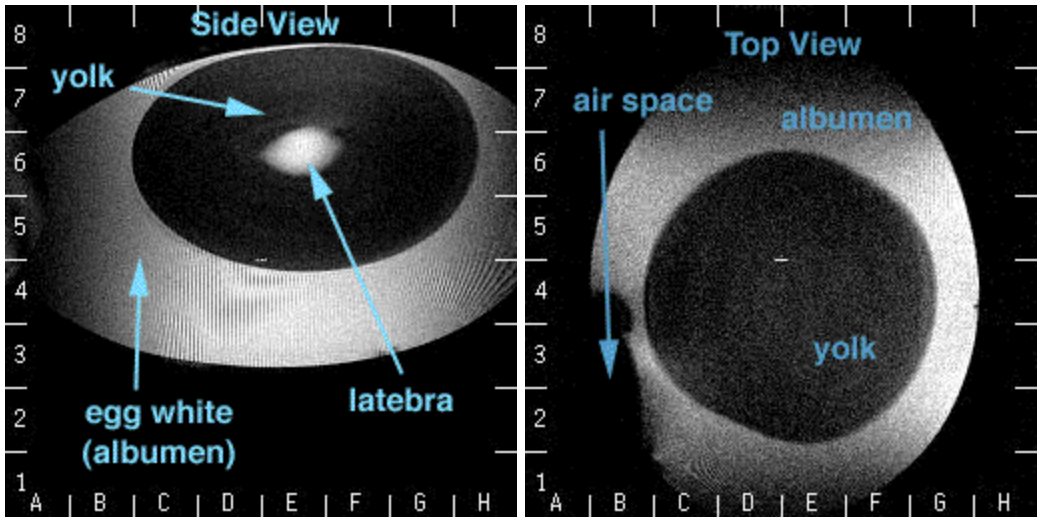
2.1.2 Chicken egg chemical and functional compositions

The average weight of chicken egg is around 58g, comprising various components including up to 11% lipids, 12% protein, and 74% water (Belitz et al., 2009). The yolk forms between 30 - 36% of the total fresh whole chicken egg weight, with dry matter content of a freshly laid yolk being between 50 to 52% depending on the age of the laying chicken (Anton, 2007). The fraction of the three major egg parts which are yolk, albumen (egg white), and shell, together with their respective cogent ingredients are as shown in Table 2.1.

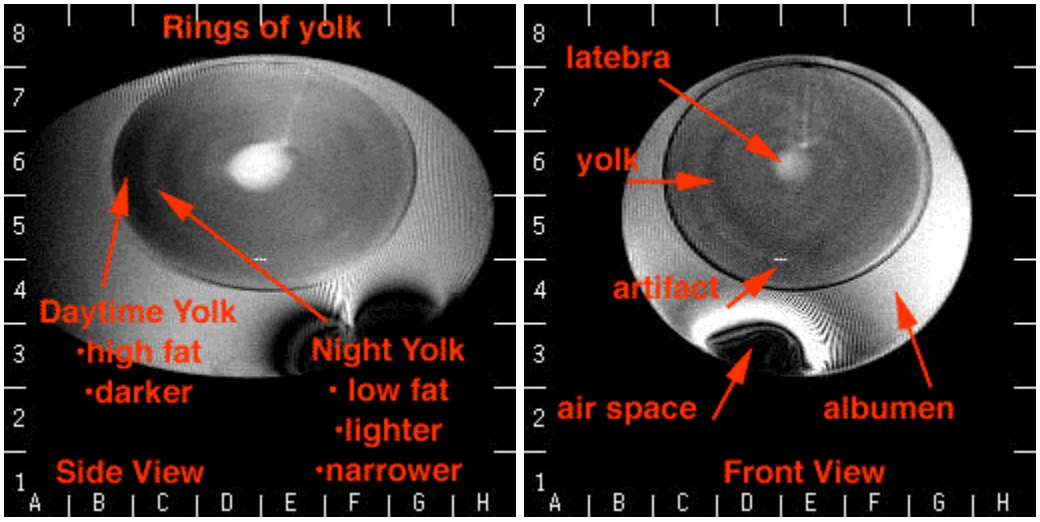
Table 2.1. Chemical composition of chicken egg (Belitz et al., 2009).

Fraction	Percent of the total weight	Dry matter (%)	Protein (%)	Fat (%)	Carbohydrates (%)	Minerals (%)
Egg yolk	32.80	51.30	16.60	32.60	1.0	1.10
Albumen	56.90	12.1	10.60	0.03	0.90	0.60
Shell	10.30	98.40	3.30	-	-	95.10

According to a scientific report by the University of Illinois at Urbana-Champaign, USA (Bruce et al., 1997; Potter et al., 1998a), the yolk during its last 7 to 8 days of development is deposited in ring like layers of white and yellow yolk. These ring-like layers though mostly not visible at earlier days of yolk formation (Figure 2.4a) is usually visible under MRI image in the last days of yolk formation as depicted in Figure 2.4b. Whereas the yellow yolk (rich in lipids or fats, and deposited during the day) appears dark in the MRI image, the white yolk (abundance in protein, and deposited during the night) is seen as narrow white bands on the MRI image (Potter et al., 1998b). The white yolk was further reported to be present directly below the nucleus (position of potential future embryo development) in the latebra and the nucleus of pander and observed to be arranged concentrically all over the yellow yolk (see Figure 2.2). The existence of



a.



b.

Source: (Potter et al., 1998a)

(chickscope.beckman.uiuc.edu/explore/embryology/day01/mri.html)

Figure 2.4 MRI images of early and late days yolk formation (a), early days of yolk formation (b), last days of yolk formation

yellow and white yolk has earlier been attested to in literatures (Okubo et al., 1997; Romanoff & Romanoff, 1949), where it was reported that the white yolk originated from the maturing white follicle in the ovary. It was also pointed out that several structures like the latebra, nucleus of pander, and embryonic disc all originated from the white yolk and that the embryonic disc in the nucleus of pander is the position for embryonic development.

The understanding of the yellow and white yolk, together with its constituents and positioning in the whole egg carries a great potential of assisting towards developing a more targeted approach for early chicken egg fertility assessment. There are 22 genetically encoded (proteinogenic or protein creating) amino acids, of which 18 present in chicken egg yolk, albumen, and whole egg are as shown in Table 2.2 (Belitz et al., 2009). Amino acids as shown in Figure 2.5, can be described structurally as organic compounds with two functional groups namely amine (-NH_2) and carboxyl (-COOH), together with a side chain -R group (specifically related to each amino acid).

Table 2.2. Amino acid composition of chicken egg in g/100g edible portion (Belitz et al., 2009).

Amino acid	Abbreviation	Egg yolk	Albumen	Whole egg
Alanine	Ala	0.82	0.65	0.71
Arginine	Arg	1.13	0.63	0.84
Asparagine	Asx	1.37	0.85	1.20
Cysteine	Cys	0.27	0.26	0.30
Glutamine	Glx	1.95	1.52	1.58
Glycine	Gly	0.57	0.40	0.45
Histidine	His	0.37	0.23	0.31
Isoleucine	Ile	1.00	0.70	0.85
Leucine	Leu	1.37	0.95	1.13
Lysine	Lys	1.07	0.65	0.68
Methionine	Met	0.42	0.42	0.40
Phenylalanine	Phe	0.72	0.69	0.74
Proline	Pro	0.72	0.41	0.54
Serine	Ser	1.31	0.75	0.92
Threonine	Thr	0.83	0.48	0.51
Tryptophan	Trp	0.24	0.16	0.21
Tyrosine	Tyr	0.76	0.45	0.55
Valine	Val	1.12	0.84	0.95

Considering the white yolk regions for in-depth analysis and targeting the protein constituents in chicken egg yolk can open new door of research opportunities towards identifying specific biomarker for differentiating between fertile and non-fertile eggs prior to incubation.

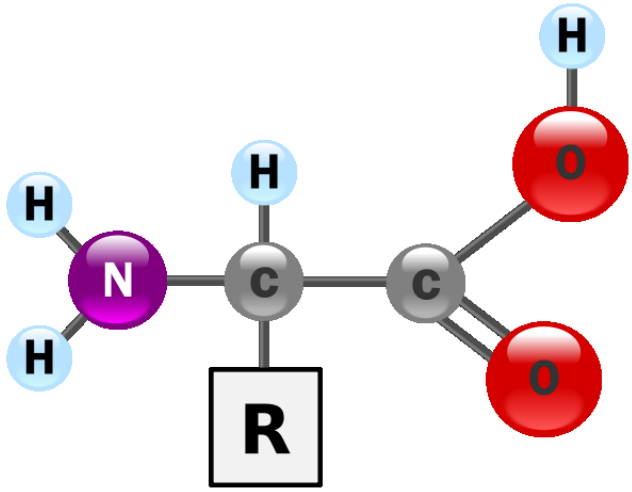


Figure 2.5 Amino acid structure showing its various functional groups
(Source: https://en.wikipedia.org/wiki/Amino_acid)

2.1.3 Chicken egg fertility and incubation

Chicken eggs are said to be fertile if the hen that laid the eggs were raised together with roosters, otherwise the eggs would be infertile. Fertilization is established from the unison of the rooster sperm with the hen's matured ovum, be it naturally or artificially through a process known as artificial insemination. While majority of the global hatchery (fertile) eggs are mostly produced from mother hen raised together with roosters, grocery store (table) eggs are from hen raised without roosters. It is only fertile eggs that carries embryonic development potentials under incubation conditions of around 55% relative humidity and temperature of 37.8^0C. Notwithstanding, it is not all the so-called fertile eggs that ends up becoming chickens under

incubation conditions, as some indeed eventually do turn out to be non-fertile eggs. Hence, there is a need to know and understand the difference between hatchery fertile and non-fertile eggs.

According to (Bakst, 2010; Wilson, 2010), fertilized eggs contain blastoderm, while unfertilized eggs contain germinal disc (blatodisc). The blastoderm (in a fresh opened egg) is visually seen as a symmetrical circular ring of about 3-4mm in diameter, having a less-dense "Area Pellucida" and denser (whitish band) "Area Opaca" regions around its perimeter (Figure 2.6a). The blastodisc in comparison to the blastoderm, has a smaller diameter (about 2.5mm), and looks like an asymmetrical solid spot, with no regional differentiation (Figure 2.6b). Under a stereomicroscope, the Area Pellucida (AP) and Area Opaca (AO) are clearly differentiated in the fertile egg blastoderm (Figure 2.6c). On the other hand, the germinal disc visualization on a stereomicroscope (Figure 2.6d) is usually characterized by many vacuoles (bubbles), anciently known as lacunae (Romanoff & Romanoff, 1949).

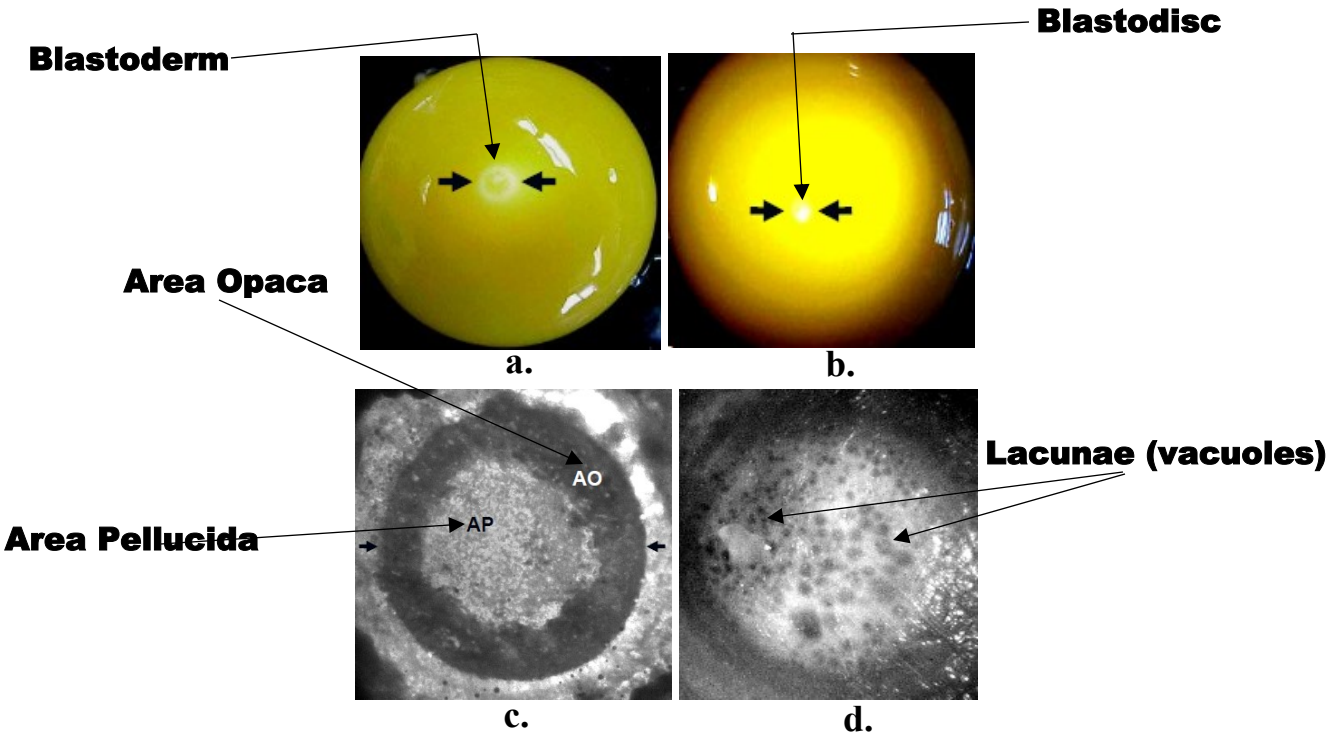


Figure 2.6 Chicken egg fertility identification: (a) and (c) – blastoderm; (b) and (d) – blastodisc
Source: (Bakst, 2010)

2.2 Industrial challenge

Complete hatchability of incubated eggs remains of a great economic concern to poultry farm owners globally. Several factors including but not limited to environmental and genetic factors were known to cause decline in chicken egg fertility and hatchability (King'Ori, 2011). While some of these factors could be arrested prior to egg production and incubation, some are only traceable after the havoc is already done. Hence, fertility and hatchability rates continue to dwindle from year to year resulting in losses of million of dollars annually. Figure 2.7 showed the total hatchery egg (both for layers and broilers) production in Canada for the years 1994 through 2017. Apart from in the year 2010 that there was a sudden jump in the number of hatched eggs, with a corresponding decline in the number of unhatched eggs; other increases in the amount of

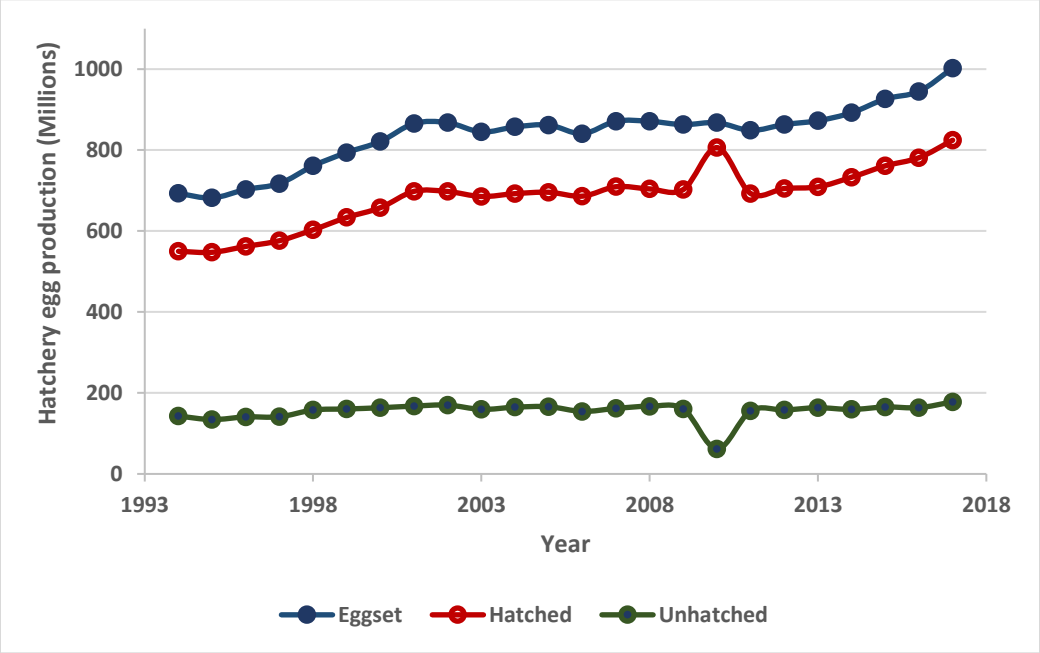


Figure 2.7 Canadian hatchery egg production from 1994 to 2017 (Data extracted from http://aimis-simia.agr.gc.ca/rp/index-eng.cfm?menupos=1.01.01&pdctc=&r=206&LANG=EN&action=pR).

hatched eggs over the years have been relative to corresponding increases in the total number of egg set available for incubation. It was therefore observed that the amount of unhatched eggs over the years have not reduced, having its least of over 134 million in the year 1995 and its peak of over 178 million in the recent past year 2017. With about 39.8 million eggs sold in the year 2017 via exportation to the United States at a sum of about 68.8 million dollars, over 300 million dollars worth of eggs were wasted as unhatched eggs in 2017. Except for year 2010, where hatchability rate stands at about 93%, Figure 2.8 showed that hatchability rates over all other years between 1994 and 2017, have not seriously improved being in the range of 79 % to 83 %. Therefore, it is very critical to determine fertility and viability of chicken eggs, prior to incubation. A fast, online and non-destructive pre-screening of eggs for fertility identification before being passed for incubation would save industries both immediate and impending losses.

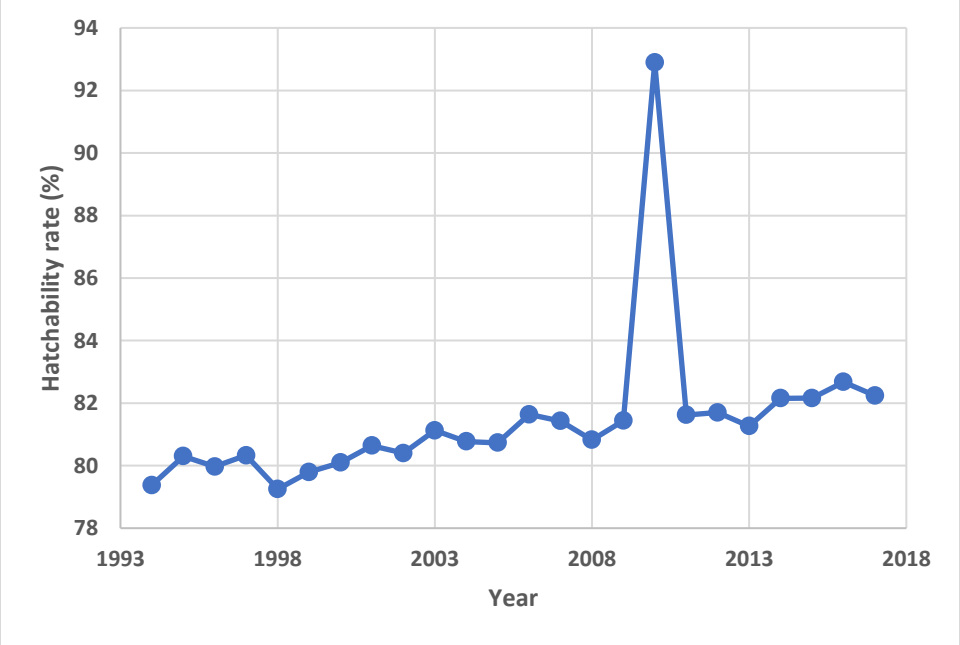


Figure 2.8 Hatchability rates for Canadian egg production (Data extracted from: http://aimis-simia.agr.gc.ca/rp/index eng.cfm?menupos=1.01.01&pdctc=&r=206&LANG=EN&action=pR).

2.3 Assessment of chicken egg fertility

2.3.1 Traditional method

The traditional method of determining fertility and separating fertile eggs from non-fertile eggs, termed candling is as depicted in Figure 2.9. Not only is this method slow and labour intensive, but also about 5% of the whole egg set are randomly candled on day 10 while remaining 95% are left to chances. In the long run, larger percentage of non-fertile eggs ends up being incubated which usually exposes the whole egg set to contamination in the case of exploder eggs. Not only that, millions of dollars ended up being lost every year as a result of these bottlenecks (Ernst et al., 2004; Lawrence et al., 2006), which include incubation space, energy, and egg wastages.

Figure 2.9 Traditional candling operation

(adapted from: https://www.eggs.ca/onthefarm/article/4/the-egg-grading-station)

2.3.2 Machine/computer vision

Great advancement has been achieved over the years for safety inspection and quality sorting of agricultural and food produce (ElMasry & Sun, 2010b). According to (Du & Sun, 2006), machine vision technology adopts image processing and analysis procedures, in combination with a set up including illumination system, and personal computer connected to a form of mechanical or electrical device. (Das & Evans, 1992a, 1992b) used machine vision to recognize fertility of hatching eggs in conjunction with histogram characterization and neural network methods. Obtained accuracies for the work were low at early days of incubation but high on days 3 and 4 incubation. These results can be attributed to the limitations of machine vision approach ranging from inability to detect intrinsic properties and to handle difficult classification tasks (Du & Sun, 2004; ElMasry & Sun, 2010b). Since various food analysis situations exist, necessitating acquiring information from inside the sample, rather than from the outside, machine vision technology might not be the best appropriate for early chicken egg fertility detection. Also, because acquisition and analysis is usually done using the three RGB spectra channels in the visible wavelength range of the electromagnetic spectrum (ElMasry & Nakauchi, 2016), it deprives users the advantage of benefiting from considering wider range of wavelength bands like those in the near, mid and far infrared regions.

2.3.3 Spectroscopy and hyperspectral imaging

Both mid-IR and NIR spectroscopy are similar based on their fundamental principles of operation, which entails consideration of molecular vibrations (Pasquini, 2003; Siesler et al., 2008). However, due to different excitation conditions of both mid-IR and NIR spectroscopy depending on the product's compounds of interest, the relationship between the functionalities of the molecules being examined and the corresponding absorption intensities differ considerably and

thus leading to a significantly different responses of the same molecular vibration, from the same applied fundamental technique (Siesler et al., 2008). Mid-IR spectroscopy involves mostly fundamental vibrations and is found in wavelength regions between 2500 and 25000 nm (4000-400 cm^{-1}) of the electromagnetic spectrum. NIR spectroscopy on the other hand entails radiations that are higher than that in mid-IR and found in the wavelength regions between 800 and 2500 nm (12500-4000 cm^{-1}).

NIR spectroscopy is among the most popular in the food industry. The absorption bands viewed in the NIR region are from overtones and combination bands of C-H, N-H, O-H, and S-H bending and stretching vibrations. Hence the NIR technologies are applicable to all organic compounds abundant in C-H bonds (like petroleum derivatives), O-H bonds (like carbohydrate, moisture, and fat), and N-H bonds (like amino acids and proteins). Since all biological substances consist of numerous amounts of O-H, N-H, and C-H molecular bonds, NIR radiation striking a sample, produces a multiplex spectrum carrying both quantitative and qualitative information about the specific sample (ElMasry & Sun, 2010b).

Visible (VIS) and Near Infrared (NIR) spectroscopy have been widely utilized in assessing internal quality of Agricultural products (Abdel-Nour et al., 2011; Giangiacomo & Dull, 1986; Williams & Norris, 1987). Norris (1996) studied the effect of storage on optical properties of shell eggs in the NIR region. Even though, progressive changes were noticed in the spectral data during storage periods, it was further observed that there was no correlation between these changes and internal egg quality indices, thus necessitating further researches and/or improvement in existing technology to make this non-destructive approach more relevant for industrial applications. (Bamelis et al., 2002) adopted a spectrophotometric method known for blood detection in Table eggs to assess early embryonic detection potential in chicken eggs. The work reported embryonic

development detection to be possible from day 5 (120 h) of incubation. This late detection might be partly related to the operational mode of spectroscopy techniques of being able to obtain only point information and so disadvantageous, should the information of interest not be present in the pixel spot measured. This disadvantage is catered for in the hyperspectral imaging technology.

Hyperspectral imaging being an improvement on conventional spectroscopy and machine vision has witnessed a wide publicity in recent times due to its combine ability to considering both spectra and spatial information from targeted samples (Kim et al., 2002; Sun, 2010). The near-infrared (NIR) regions in particular have been successfully used for food (including chicken egg) quality and safety analysis (Abdel-Nour et al., 2011; Chen, 1993; Chen et al., 1998; Osborne et al., 1993; Williams & Norris, 1987), and could as well proof effective for chicken egg fertility assessment studies. Hyperspectral imaging, though a relatively new and emerging technology has proven more advantageous than spectroscopy and computer vison due to its chemical free assessment, non-destructive, and non-invasive nature, spatial distribution visualization, fast image acquisition potentials, little or no sample preparation, eventual simple and fast analysis method, flexibility in region of interest (ROI) selection (ElMasry & Sun, 2010b; Liu & Ngadi, 2013; Yanenko & Velikanov, 2014), and ability to handle sample heterogeneity.

2.4 Hyperspectral imaging technology and Instrumentation

Optical sensors have been known to provide great potentials for non-destructive analysis of agricultural products. Imaging and spectroscopic techniques have been widely studied and used in various field of endeavours including agricultural applications (Bamelis et al., 2002; Steiner et al., 2011). However, spectroscopy and conventional imaging approaches are limited when it comes to obtaining adequate information from individual food items (Qin, 2010). Due to recent advancement in imaging and spectroscopy technologies, hyperspectral imaging has emerged as a

preferable alternative for quality assessment and safety control of agricultural produce (Ariana & Lu, 2008; Bodkin, 1997; Bodkin et al., 2005; Qin, 2010). Generally, a hyperspectral system comprises of a light source, a wavelength dispersion device, and an area detector. Figure 2.10 showed a schematic description of the emergence of hyperspectral imaging system from conventional imaging and spectroscopy (Qin, 2010). The choice of illumination source is a very

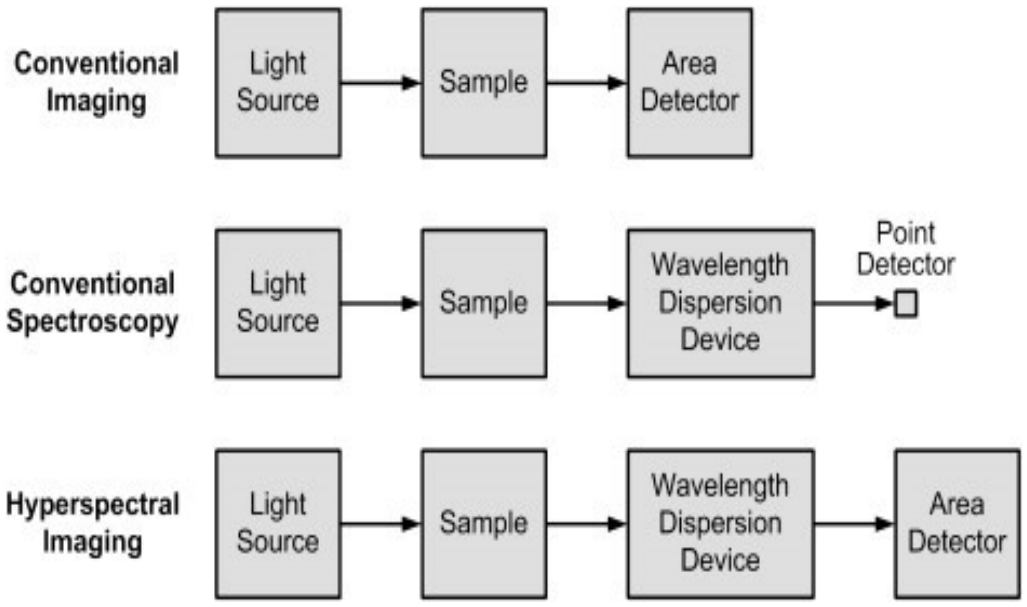


Fig. 2.10 General system configurations for conventional imaging, conventional spectroscopy, and hyperspectral imaging (Qin, 2010).

critical factor of consideration in the planning and setting up of any imaging system. For example, the nature of the emitted spectrum and the amount of light intensity reaching the object of interest will influence the subsequent quantity of light absorbed, transmitted or reflected from the object. Hence, only the wattage rating of lamps is not enough in selecting a suitable light source for imaging application but the illuminance of the said light source. Some lamps with higher wattage ratings have been shown to have lower illuminance when compared to other lamps of lower

wattage ratings. This is because larger percentage of light emanating out of some lamps is being lost as heat and so unusable. According to Qin (2010), halogen lamps are the most popular broadband illumination sources that have been successfully used in the visible (VIS) and near-infrared (NIR) wavelength regions for hyperspectral imaging. Specific applications include but not limited to "pits detection in tart cherries", "bone fragment detection in chicken breast fillets" and "detecting fertility and early embryonic development of chicken eggs" (Liu & Ngadi, 2013; Qin & Lu, 2005; Yoon et al., 2008). Other illumination sources with the potential of gaining wide acceptability in the nearest future are lasers, tunable sources and light emitting diodes (Brauns & Dyer, 2006; Chao et al., 2007; Jestel et al., 1998; Klein et al., 2008; Lawrence et al., 2007; Mueller-Mach et al., 2002; Noh & Lu, 2007; Wabuyele et al., 2005).

2.4.1 Principle of operation

Hyperspectral imaging works on the optical principle of light and its interaction with matter. When light energy (photon) falls on an object, you do not see the light, but the amount of light energy available determines how much of the object you see based on the influence of the light. The brain behind hyperspectral imaging system as a tool for non-destructive food analysis is therefore based on the understanding of light photons interaction with the molecular structures of food samples (ElMasry et al., 2012; ElMasry & Sun, 2010b). As light energy strikes an object, the incident light reaching and interacting with the object can be reflected, absorbed or transmitted as depicted in Figure 2.11. These reflected, absorbed or transmitted light carries important information from the passing medium (object) and can be used for both qualitative and quantitative predictions. Figure 2.12 showed the summarized basic steps involved in hyperspectral imaging analysis, with an overall intention of being translated into a multispectral imaging system, which is usually more economically built and suitable for an online and real-time industrial application.

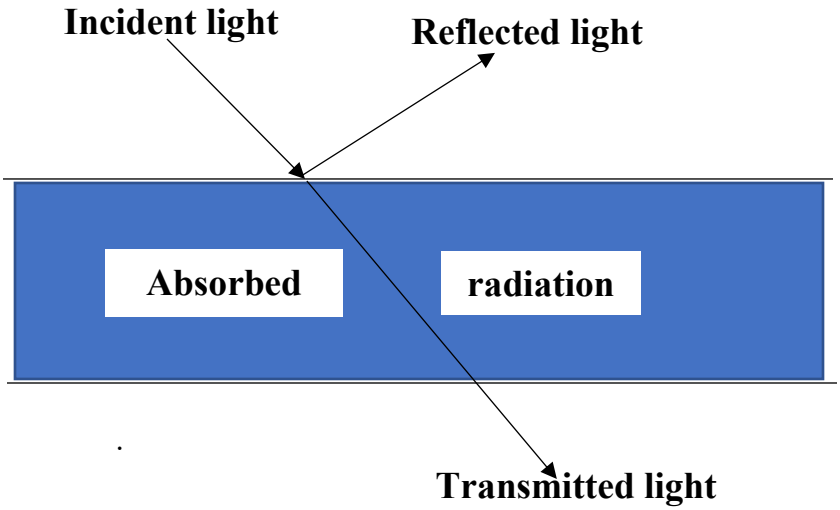


Fig. 2.11 Hyperspectral imaging principle of operation

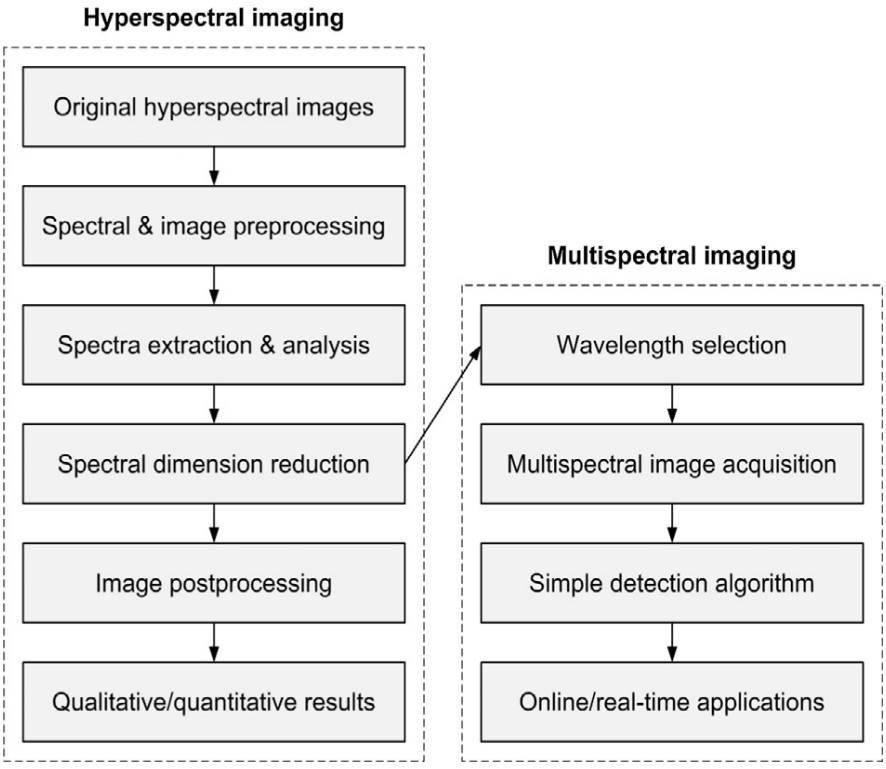


Fig. 2.12 Hyperspectral-Multispectral imaging flow chart (Qin et al., 2013).

2.4.2 Image acquisition, data extraction, and spectra pre-processing

Knowing well that it is practically impossible to get any useful information from less qualitative data, obtaining a high-quality image therefore becomes very critical in hyperspectral imaging and related researches. Consistency and accuracy are needed in various settings including acquisition mode, illumination type and arrangement, detector selection, spectral and spatial resolutions, frame rate, scanning speed, camera exposure/integration time, and calibration (ElMasry & Nakauchi, 2016; ElMasry et al., 2013; Lewis et al., 2007). After image acquisition, spectral information (X-matrix) is usually extracted from a segmented region(s) of interest (ROI's). These ROI's stand for expected or actual locations of targeted biological or quality attributes in the acquired image (Kamruzzaman et al., 2013; Sone et al., 2012). A corresponding response or reference information (Y-matrix) is also collected following standard conventional (usually destructive) method. The response information should ideally be obtained at the exact ROI's from which the spectra (X-matrix) data has been collected (ElMasry & Nakauchi, 2016). Depending on the degree of quality achieved in the acquisition/extraction stage, hyperspectral images/data usually contain noise and some unwanted variabilities due to various other factors including but not limited to detector anomalies, particle size variations and light scattering effects. Hence, spectra pre-processing is usually implemented to minimize the effect of the above-mentioned problems. Spectral pre-processing can consist of one or combination of the following techniques namely: filtering, smoothing, normalisation, mean centering, scaling, standard normal variate (SNV), multiplicative scatter correction (MSC), orthogonal signal correction (OSC), derivatives, detrend, Fourier and Wavelet transforms (Esquerre et al., 2012; Vidal & Amigo, 2012).

Filtering is purposely employed to remove non-informative variables in a data set. This approach eventually results in improved statistical power during a downstream multivariate analysis. Smoothing filters on the other hand are usually implemented to reduce noise, while simultaneously preserving the number of variables (Hackstadt & Hess, 2009; Xia & Wishart, 2016). Normalization is used in adjusting samples to approximately the same scale. The most common approach is the use of mean centering (dividing each instance of a data matrix by its mean) or autoscaling (mean centering + division by standard deviation of individual variable). Other forms of normalization like normalization by sum, median, reference sample, reference feature and data transformation such as logarithmic and cube root transformations are used for general purpose modification for variability among instances and to make individual attributes more comparable (Camo, 2018; Xia & Wishart, 2016). According to (Camo, 2018), MSC and SNV are used to adjust for multiplicative and /or additive effects in spectra data, including removing particle size effect and correcting for path length variation. Detrending, used for removing nonlinear drifts in spectroscopic data is also versatile in reducing data baseline shift, curvature and multicollinearity, when implemented in conjunction with SNV. Derivatives of various orders are row-oriented transformation widely used to reveal hidden information that might not be visible considering the raw data spectrum. Most of the pre-processing procedures as iterated above are commonly implemented with the aid of specialised software packages including Unscrambler, Matlab, WEKA, SAS, JMP, MetaboAnalyst, and other related packages.

2.4.3 Dimensional reduction techniques

Since HSI sensors produce a great number of spectral bands, the major task in HSI analysis entails dealing with a very huge amount of data. This does not only increase computational difficulties but also in the long run affects severely classification accuracies. Hence, various feature

extraction techniques are usually employed to reduce dimensionality and extract important features from HSI data and thereby eliminating as much as possible spectra redundancy from the acquired multidimensional HSI data (Renard et al., 2008). The most widely used dimensionality reduction techniques are the linear methods of principal component analysis (PCA) and multidimensional scaling (MDS). Others include but not limited to partial least square (PLS), ISOMAP and Autoencoder (Cox & Cox, 2000; Partridge & Calvo, 1998). Some of the existing dimensional reduction techniques can select important features and simultaneously extract new features for discrimination, and this is the reason some researchers do mistakenly accept feature extraction and feature selection to be the same. There exist indeed other specific approaches solely for feature selection. Whereas feature extraction creates new attributes (transformed features) from functions of the original features, feature selection chooses a small subset of the original attribute set, that performs optimally under some criterion function. Feature selection has been widely accepted as appropriate for hyperspectral imaging data due to its aiding speedy information acquisition/processing and eventually resulting in great cost savings (Nakariyakul, 2007).

During feature selection, the choice of subsets per time to be learned are usually determined using the embedded, filter and /or wrapper methods (Ladha & Deepa, 2011; Saeys et al., 2007). While the filter approach considers the appropriateness of the selected features, independent of the classifying algorithm, the wrapper method requires a classifier to evaluate feature appropriateness, but also can be computationally burdensome. Whereas the filter techniques are classifier independent, simple and fast; they are limited due to their dark knowledge of the interaction between feature subset search and classifier (Liu et al., 2014). This disadvantage with the filter techniques is catered for in the wrapper and embedded methods. Also, there exist multivariate filters purposely developed to overcome limitation of the conventional filter approach, and these

include the information gain, correlation and learner-based feature selection techniques (Hall, 1999; Jason, 2016; Liu et al., 2014). Whether it be filter, wrapper, or embedded based feature selection system, their implementation is always in conjunction with various search algorithms. Such algorithms according to the work of (Nakariyakul, 2007) have been described in terms of optimal, quasi-optimal, and ratio feature selection algorithms.

A feature selection algorithm is said to be optimal if it chooses the best subset of "m" of "n" attributes, with the best "m" determined as being the subset having maximum value for a chosen criterion function. While the optimal algorithms include the exhaustive search and the branch-and-bound (BB) algorithms, quasi optimal algorithms include the sequential forward floating (SFFS) and sequential backward floating (SBFS) selections (Ferri et al., 1994; Kudo & Sklansky, 2000; Pudil et al., 1994). The choice of selection algorithm is greatly dependent on the number of "m" subsets to be evaluated per time for criterion function computation. The greater the number of criterion functions to be computed, the greater the search time will be, and this is invariably dependent on the total number of original "n" features. For example, an exhaustive search for four best wavelength band attributes from a 400-featured data set will need a search of "400 *combination* 4" ($^{400}C_4$) ≈ one billion subsets (Nakariyakul, 2007). This search is known to be exponentially increased, should there be consideration of ratio features.

2.4.3.1 Ratio feature selection algorithms

Ratio features have been reported of having better discrimination potentials than individual features. (Guyon & Elisseeff, 2003) indeed confirmed experimentally that an attribute that is completely useless alone, can provide notable performance improvement when considered alongside other attributes. It was also buttressed in the same work that two features that are redundant by themselves can be useful together. Hence ratio feature consideration has began to

gain increasing interests in feature selection approaches. (Xia & Wishart, 2016) reported the use of a PLSDA learner and ranking algorithm to select up to 100 ratio features, in an online metabolomic analysis platform. Ideally, depending on the number of original "n" attributes, there could be numerous ratio features to be computed. For example, there would be "167 *combination* 2" ($^{167}C_2$) = 13, 861 possible ratio attributes out of 167 original wavelength attributes. To choose only two best sets of ratio attributes from the above will require exhaustive search to calculate criterion functions for all ($^{13,681}C_2$) ≈ 93 million combinations of two sets of ratio features. Since these number of combinations can greatly increase exponentially depending on original "n" attributes and "m" subsets of ratio features needed, quasi optimal algorithms including SFFS and SBFS have been suggested feasible for large ratio features computation and eventual ratio subset selections. (Nakariyakul, 2007) suggested a new adaptive branch and bound (ABB) algorithm, an improved sequential forward floating selection (ISFFS) algorithm, and a fast ratio feature selection algorithm, which were all successfully tested on chicken skin tumor and chicken contaminant hyperspectral imaging data. Even though there have not been many works on using ratio features in chicken egg fertility studies, (Bamelis et al., 2002) reportedly used the 527/610 nm ratio popularly used in commercial blood detectors for early embryonic development detection in chicken eggs. The work however concluded that embryonic development detection with visible light transmission is not directly correlated with blood formation. Therefore, there is a need for more studies to consider many possible ratio features from chicken egg fertility data, before arriving at optimal ratio feature selections.

2.4.4 Multivariate Analysis (Post Processing)

Multivariate analysis (MVA) can be categorised into three major areas namely: exploratory data analysis (EDA), regression analysis and discrimination analysis (classification). EDA is also

often called data mining and it is the common approach used towards understanding deeper insights into large and complicated data sets. While regression analysis assists in model development towards prediction of new and future events, classification on the other hand is a versatile tool useful in research, development and market analysis, towards handling categorical data. Even though each method of MVA used on its own can produce worthwhile results, effective combination of these methods can bring about outstanding revelations about the system under study (Swarbrick, 2012). Two main approaches commonly employed in EDA are cluster analysis and principal component analysis. Whereas cluster analysis achieves the job of isolating objects into groups (clusters) in which members of an identified cluster are related to each other, PCA analyses variability in data set, thereby understanding correlations between samples and variables.

2.4.4.1 Regression analysis and predictive modelling

Regression models are models used to predict numeric outcome and are therefore also called quantitative models (Kuhn & Johnson, 2013). Regression analysis produces only continuous responses. According to Swarbrick (2012), regression analysis often involves two data sets comprising of the predictors (independent) and dependent (response) variables. Independent variables are already known measurements to make model from, towards predicting the required output. Dependent variables on the other hand are the responses being modelled from the predictors. Responses depend greatly on the predictors used in the model. Widely used multivariate regression methods include multiple linear regression (MLR), principal component regression (PCR), and partial least squares regression (Amigo et al., 2009; ElMasry & Nakauchi, 2016).

2.4.4.2 Discrimination analysis/classification algorithms

Discrimination is the term used to describe separation (or division) of a group of samples into one or more classes based on characteristic features in the samples. Discrimination

(classification) has been known to be a very crucial task in pattern recognition. Discriminative models are also known as qualitative or categorical response models. Two basic approaches to solving classification problems in general are those of unsupervised and supervised algorithm techniques. In unsupervised learning, data are grouped based on some similarities/dissimilarities or characteristics inherent in the data set and analyst may not have a priori knowledge about the grouping. Supervised learning on the other hand gives the opportunity of having the idea of what factors, input or predictors that will have impact on the output response even though one might not have the complete understanding of the relationship between the response and the predictors. Notable methods used in unsupervised classification include K-means, K-medians, hierarchical cluster analysis and principal component analysis. For supervised classification, the following techniques are often employed: soft independent modelling of class analogy (SIMCA) with PCA, K-nearest neighbours (KNN), linear discriminant analysis (LDA), Logistic regression, partial least squares discriminant analysis (PLSDA) and support vector machine classification (Swarbrick, 2012). Other unsupervised and supervised learning algorithms which are independent component analysis (IDA), Fuzzy one class Support vector machines (FOCSVM), and associative classification have also been described elsewhere (Dong et al., 1999; Li et al., 2004; Naik & Kumar, 2011; Yu et al., 2011). It must be pointed out that for classification tasks, non-supervised learning approaches are not definitive in implementation but usually inform analyst of possible need of progressing into a more conclusive supervised learning methodology or peradventure stop moving ahead if the non-supervised learning results were deemed unsatisfactory. For any task with the end goal of classification (and not exploration), non-supervised learning technique might not stand alone unless used in conjunction with a supervised learning approach. However, supervised

learning methodology is standard for classification task and could stand alone towards a conclusive analysis for a discrimination problem.

2.4.5 Performance evaluation

Regression analysis performance are usually assessed in terms of the following criteria namely: correlation coefficient of calibration and validation (R_c and R_v), coefficient of determination (R^2), root mean square error of calibration and validation ($RMSE_c$ and $RMSE_v$) and predicted sum of squares (PRESS). While RMSE is a function of the model residuals, R^2 can be understood as the fraction of the data information being explained by the model and it is usually in close relation to correlation (Kuhn & Johnson, 2013). For classification, performance is usually evaluated in terms of the overall accuracy (OVA), which shows the overall percentage of correctly classified instances as against incorrectly classified instances. This criterion has however been regarded as misleading when considering data set of an imbalanced nature (Nguyen et al., 2009), thereby leading to the choice of the confusion matrix evaluation criteria. Details of the confusion matrix criteria and other related evaluation metrics for an imbalanced data scenario are as described in section 2.7.

2.5 Hyperspectral imaging for chicken egg fertility assessment

There have not been many works done on using hyperspectral imaging for chicken egg fertility assessment. The frequently occurring five published studies have been carried out by only three notable research groups in the USA, Canada, and China (Lawrence et al., 2006; Liu & Ngadi, 2013; Smith et al., 2008; Smith et al., 2005; Zhang et al., 2014). In the work of (Smith et al., 2005), early fertility and embryonic development detection of hatching eggs were assessed in two separate experimental settings, using ratio wavelength 576/655 nm and ratio ranges between 576 nm and 682 ± 13 nm, for both brown and white eggs. While 12 eggs were imaged daily in 2 replicates, and

without replacement for 4 incubation periods in the first experimental set up for white eggs, 12 eggs were imaged daily in 2 replicates, and with replacement for brown eggs in the second experimental set up. The experiment 1 outcome reported 1 of 46 fertile eggs detected for total eggs on days 0 and 1 incubation, 60% fertile and early embryonic development detection on day 2, and 91% fertility accuracy on day 3. Experiment 2 confirmed all considered eggs to be fertile upon break out analysis and so fertility classification accuracy was impossible to be tested at this point. Apart from the fact that this work did not solve the early discrimination problem prior to incubation, the sample size was small, the replacement and non-replacement approaches for different types of eggs did not give a good basis for comparing performance of white and brown eggs, results validation was not done or reported, and it was not clear or stated explicitly the classification algorithm used in the study.

A follow up hyperspectral imaging study similar to that of (Smith et al., 2005) was conducted by (Lawrence et al., 2006) on brown shelled eggs in visible transmission wavelength regions of 420 to 840 nm. Egg samples remain relatively the same as in previous study, but Mahalanobis Distance Classification and PLSR algorithms were used for models' development. Preliminary data analysis was carried out on both spectral and spatial information. Also, preprocessing operations of smoothening and multiplicative scatter correction were implemented. In the same vein, results were validated using leave one-out cross validation (LOV). There was no observed improvement in spectral models, using textural and morphological attributes. Classification results of 100% on days 2 and 3, 95.8% on day 0, and 91.7% on day 1 were reported for Mahalanobis Distance (MD) Classifier. PLSR modelling algorithm on the other hand achieved 100% accuracy for all days of incubation, with LOV. Despite the presented results looks optimistic, break out analysis showed there were no non-fertile eggs present in the sample size

considered and hence the reported results were best regarded as being for egg embryonic development and not fertility. Therefore, fertility recognition problem prior to incubation at this juncture remains unsolved. It was also seen here the difficulty of having non-fertile eggs presence for training, in a small sample size collection. Future robust models will indeed need large data size for calibration, validation and testing.

In another subsequent study by (Smith et al., 2008), same 12 hatchery fertile eggs were used but now with a matching up 12 non-fertile eggs acquired from flock raised without roosters. Hyperspectral images were then collected in 8 replicates in the visible wavelength regions of 400 to 1000 nm. Adopting a MD Classifier, in conjunction with PCA, 5 replicates data were used for calibration and remaining 3 replicates used for validation. New set of 3 replications of 30 eggs each (of randomly mixed fertile and non-fertile eggs) were also reportedly used for verification. The outcome of this study presented overall accuracy for validation set of eggs as 71% on day 0, 63% on day 1, 65% on day 2, and 83% on day 3, with lower verification results reported as 51% on the average. This study seems to be the first standard procedural set up to handle chicken egg fertility problem, using the non-destructive hyperspectral imaging technology. Notwithstanding, the sample size remains inadequate, and the study concluded from obtained results that the PCA/MD classifier adopted was inappropriate for early embryonic development and fertility detection. There is therefore a need for more appropriate modelling technique(s) to capture and learn accurately information from chicken egg fertility hyperspectral imaging data.

Liu and Ngadi (2013) introduced the use of hyperspectral imaging technique in the near infrared (NIR) wavelength regions (900-1700 nm) to detect fertility and embryonic development in 174 white leghorn eggs. Mean spectral and image textural characteristics extracted using a Gabor filter algorithm were further analysed in the study. The work finally implemented

dimensional reduction technique of PCA in conjunction with K-means clustering algorithm, towards model development. Best over all classification accuracies reported were 100% on day 0, 78.8% on day 1, 74.1% on day 2, 81.8% on day 3, and 84.1% on day 4. Hyperspectral imaging potential of determining fertility prior to incubation was therefore established in this study. Owing to the knowledge that improved assessment of relevant chemical and other quality attributes are more readily obtained in the NIR region, the success of Liu and Ngadi (2013) have been linked to higher wavelength consideration and improved HSI technology (including dimensional reduction techniques of PCA and feature extraction). Despite the great promising results obtained from this study, the modelling approaches adopted were non-supervised learning techniques, and so need further confirmation using standard supervised learning algorithm(s). Whereas some non-supervised learning approaches like the PCA, k-means, k-medians, and hierarchical cluster analysis are excellent with identifying/understanding grouping and clustering patterns in multidimensional data, they are limited when the end target is discrimination (Barker & Rayens, 2003). Furthermore, unsupervised classification is always the starting point in any discrimination problem and should necessarily be followed by supervised classification (Swarbrick, 2012), towards an industrial adoptability consideration.

In similarity to earlier works, (Zhang et al., 2014) compared the spectral and image morphological attributes of 90 green shelled chicken egg towards hatchability detection. The study used a single selected optimum wavelength of 822 nm out of the considered visible wavelength regions between 400 to 1000 nm, and PCA in conjunction with Learning Vector Quantization Neural Network (LVQNN) were adopted as modelling algorithms. Overall accuracies using spectral characteristics were reported as 65% on day 0, 63% on day 1, 60% on day 2, 77% on day 3, and 83% on day 4. Results accuracy using morphological attributes were 72% on day 0, 70%

on day 1, 76% on day 2, 97% on day 3, and 100% on day 4. The worthwhile improvement in accuracy brought about by using morphological attributes was only possible on days 3 and 4 incubation. Therefore, this study also did not solve the problem of fertility detection prior to incubation. The outcome of this study is however consistent with the earlier report of (Lawrence et al., 2006) that no textural or morphological attributes considered in the visible wavelength regions for brown eggs brought any significant improvement to models built from only spectral data.

From all the cases considered, it was only the work of (Liu & Ngadi, 2013) that showed the greatest potential of using HSI technique for early fertility discrimination, especially prior to incubation, and this was possible considering the NIR wavelength regions of the light spectrum. It was clear from the reviews that further works towards building futuristic robust model for chicken egg fertility early detection will need to take care of lack of enough data, rare class data acquisition problem (too little or non-availability of non-fertile eggs for learning), and appropriate analysis/modelling techniques.

2.6 Imbalanced data problem in Agricultural and food processing applications

A data set is said to be imbalanced if the classification groups in the data are not equally represented (Chawla, 2009). The specific group with very few training examples is usually called the rare (minority or positive) class, while the other with many examples is called the prevalent (majority or negative) class. Imbalanced data situation exists in most fields of endeavour like the biomedical, surveillance and security industries, insurance, management/finance (Artís et al., 2002; He & Garcia, 2009) and the Agricultural sectors. Due to the fact that rare cases occur infrequently, classification rules that detect small groups tends to be scarce and samples belonging to small classes are largely misclassified than those of prevalent classes (Sun et al., 2009). It is

therefore of great necessity that emerging researches in Food and Agricultural applications pay close attention towards addressing the menace of imbalance data distribution, which has long been neglected in the food and Agricultural data analyses.

Existing cases of imbalanced data scenario in the Agro-food sector include but not limited to crop-food disease and stress detection (Dale et al., 2013; Del Fiore et al., 2010; Zhang et al., 2003), fruit bruise detection (Ariana & Lu, 2008; Ariana & Lu, 2010a, 2010b; Wang & ElMasry, 2010), infectious fruit/vegetable prediction (Senthilkumar et al., 2016a; Senthilkumar et al., 2012; Senthilkumar et al., 2016b), and chicken egg fertility assessment (Das & Evans, 1992a, 1992b; Lawrence et al., 2006; Liu & Ngadi, 2013; Smith et al., 2008; Smith et al., 2005). In each of the Agro-food cases listed above, the nature of the imbalance is that of the majority class being of uttermost recognition importance as against most other external fields in which the minority class is always of uttermost recognition importance. For example, whole (unbruised) food products are ideally more abundant than bruised food products and correctly identifying all unbruised fruits/vegetables is more beneficial and economical to the food industries than misclassifying some bruised fruits. The situation is quite different with other fields such as the biomedicals, in which correctly identifying the rare class disease subjects is more critical than misclassifying some healthy control subjects. This difference in the class of uttermost recognition importance between the Agro-food cases and cases in other sectors, is a major point of consideration during analysis of Agro-food imbalanced data.

To the best of our knowledge, there have not been reported research efforts geared towards solving the imbalanced data situation in the Agricultural sector especially in food analysis research. Researchers in this field have been carrying out analysis and validating models on assumed balanced data, the results of which when applied to the real-world situation are very prone

to doubt (Kuhn & Johnson, 2016). This might indeed be the major reason for the low acceptability and adoptability of such models in real industrial settings. The end goal of this section therefore is to eventually examine the existing approaches for handling imbalanced data problem in other fields of endeavour like the biomedicals and computer sciences, with a view of adopting such approaches to the Agricultural and food processing applications.

2.6.1 Handling imbalanced data problem with chicken egg fertility classification

Having earlier identified three major areas to focus on towards building a futuristic robust model for chicken egg fertility early detection, this sub-section seeks to review possible ways of tackling the identified bottlenecks namely sample size, analysis/modelling techniques, and the rare class data acquisition problem. Solving the sample size and analysis/modelling technique challenges can be simple and straight forward by sacrificing the time, financial and human resources to acquire large enough quality data and trying such data painstakingly on pools of available classification algorithms and modelling techniques. However, solving rare class data problem is somehow complicated and very critical, since omitting it would eventually render ineffective the other solutions to sample size and modelling techniques. Handling the rare class data problem therefore takes the priority among others.

In real-world situation, chicken egg fertility/early embryonic development detection study belongs to the category of imbalanced data distribution during analysis. This is because the occurrence of fertile eggs is much more frequent than that of the non-fertile eggs in any available egg set. Indeed, it is commonly observed in the industrial settings and commercial hatcheries that only up to 10% non-fertile eggs exist in any whole egg set batch. This occurrence has brought major setbacks on the classification accuracies of most existing learning algorithms. Even though, various classification learning algorithms like the backpropagation neural network, decision tree,

nearest neighbor, support vector machine, Bayesian network, etc. have been successfully applied in many application domains; data set of imbalanced distribution still continue to be a critical bottleneck for most classifier learning algorithms (Chawla et al., 2002; Fawcett & Provost, 1997; Kubat et al., 1998; Schapire, 2003; Sun et al., 2009). Researchers in various application areas including the biomedicals and computer sciences have proposed some solution approaches which would be worthwhile to try out on chicken egg fertility and other Agro-food related researches. Such approaches are the feature selection based, data preprocessing (resampling), recognition based, cost-sensitive learning and the ensemble methods (Elkan, 2001; Nguyen et al., 2009; Phoungphol, 2013; Rokach, 2010; Seiffert et al., 2010).

2.6.1.1 Feature selection

There have been some research efforts reported on using features selection to tackle imbalanced data problem (Forman, 2003; Zheng et al., 2004). Features are usually ranked independent of their relationship with other features, thereby showing the effectiveness of individual feature in predicting the category of each sample (Phoungphol, 2013). Lê Cao,Bonnet, et al. (2009) adopted an optimal feature weighting (OFW) algorithm to select optimised features from high dimensional and imbalanced microarray data. Likewise, (Wasikowski & Chen, 2010) developed a new feature selection (FAST) algorithm based on AUC and threshold moving technique, to tackle small sample imbalanced data sets.

2.6.1.2 Data preprocessing (resampling)

Data preprocessing is otherwise known as data sampling or resampling. The approach focuses on modifying class distribution towards handling class imbalance. The technique has a major advantage of being implemented independently of any underlying classifier (López et al., 2013). The main task with the approach is to preprocess training data to minimize any divergence

between the classes, thereby improving the initial data distributions of the prevalent and non-prevalent class to achieve a more uniform number of occurrences in each class (Liao, 2008; Nguyen et al., 2009). Data preprocessing approach has been widely discussed under the following categories namely: over-sampling, under-sampling and hybrid of over-sampling and under-sampling. Liao (2008) successfully used data preprocessing methods of over-sampling, under-sampling and the hybrid of the two to classify weld flaws with imbalanced class data sets.

Over-sampling increases the number of the rare class occurrences by duplicating them until they are at pal with the prevalent class occurrences. This approach is advantageous in that all information from the majority class is kept intact and all the occurrences of the rare class are also fully considered. Notwithstanding, researchers have reported the likelihood of occurring overfitting with this method as existing copies of instances are usually exactly duplicated. Due to this set back, more sophisticated approaches have been proposed among which the "Synthetic Minority Oversampling Technique" (SMOTE) has become popular. (Chawla et al., 2002; Chawla et al., 2004; Japkowicz & Stephen, 2002). The main principle behind SMOTE implementation is to create new rare class (synthetic) examples via interpolation of various non-prevalent class instances (nearest neighbours) lying together, for oversampling the calibration data set (López et al., 2013). Due to the possible challenge of overgeneralisation largely related to the manner of synthetic samples generation, there are exist some adaptive sampling methods, proposed to reduce overgeneralisation tendencies with SMOTE implementation. These methods include the use of Boarderline-SMOTE, SPIDER2, Adaptive Synthetic Sampling, and Safe-Level-SMOTE algorithms (Bunkhumpornpat et al., 2009; Han et al., 2005; He et al., 2008; Stefanowski & Wilk, 2008).

In under-sampling, the majority class instances are reduced to a smaller set comparable to the minority class and thereby at the same time preserving all the minority class occurrences. This technique nonetheless has a drawback of loosing cogent information from the majority class occurrences and thereby degrading classifier effectiveness. However, since mode of operation with under-sampling is mostly based on data cleaning techniques, some data cleaning algorithms have been proposed to uplift the results of conventional under-sampling implementation. Such data cleaning algorithms include the Wilson's edited nearest neighbour (ENN) rule, the one-sided selection (OSS), Tomek Links, the neighbourhood cleaning rule, and NearMiss-2 method (Kubat & Matwin, 1997; Laurikkala, 2001; Mani & Zhang, 2003; Tomek, 1976; Wilson, 1972). Combination of data cleaning and resampling techniques have also been reported to have potentials of reducing overlapping commonly introduced by adopting resampling method alone, and by so doing, a best percentage of implementing both under-sampling and oversampling could be ascertained (Batista et al., 2004; Chawla et al., 2008). Furthermore, some cluster-based sampling algorithms have been reported to be useful for pre-processing before implementing under-sampling and/or oversampling (Bunkhumpornpat et al., 2012; Cohen et al., 2006; Jo & Japkowicz, 2004; Yen & Lee, 2006, 2009; Yoon & Kwek, 2005, 2007). In the same vein, the application of particle swarm optimisation or genetic algorithms for correct identification of useful examples have been shown to be very helpful with imbalanced learning (García & Herrera, 2009; Yang et al., 2009).

2.6.1.3 Recognition-based approach

This is also known as one-class learning approach. Some machine learning algorithms including but not limited to fuzzy classifiers, decision trees, neural networks and support vector machines are prone to identifying the majority class occurrences having been trained to obtain the

overall accuracy, to which the rare class contribution is but minimal. A one-class or recognition-based approach therefore offers a solution in which the classifier is modelled on the examples of the non-prevalent class (rare class) not considering the examples from the prevalent class. This approach is particularly useful when instances from target class are scarce or hard to obtain (Nguyen et al., 2009; Phoungphol, 2013). Recognition based approach has been reportedly applied in conjunction with autoencoder-based classifiers, neural networks, ensemble classifiers and SVMs (Eavis & Japkowicz, 2000; Japkowicz et al., 1995; Raskutti & Kowalczyk, 2004; Spinosa & de Carvalho, 2005). Yu et al. (2011) used a fuzzy one-class SVM on imbalanced data to detect fall in a smart room. Likewise, Manevitz and Yousef (2002, 2007) reported the successful use of one-class learning approach in document classification based on SVMs and autoencoder respectively. Unlike the conventional SVM, one-class SVM identifies instances from one group instead of differentiating all instances (Phoungphol, 2013). While considering an imbalanced genomic data set, (Raskutti & Kowalczyk, 2004) showed that one-class SVMs outperform the conventional binary-class SVMs. The study further reported that one-class learning is specifically advantageous when used in a highly dimensional, exceptionally imbalanced and noisy feature data space. Notwithstanding, (Nguyen et al., 2009) reported a notable setback with recognition-based approach learning as being its inability to handle numerous machine learning algorithms like the Naïve Bayes, associative classifications and even decision trees simply because these classifiers are built from samples of more than one-class.

2.6.1.4 Cost-sensitive learning

This is employed in practical situations where the misclassification costs are also paramount and not only the data distribution skewness. Majority of the traditional learning algorithms tend to disregard the difference between types of misclassification errors by assuming

all misclassification errors cost exactly the same. Cost-sensitive learning methods build on the merit of the fact that it is less expensive to misclassify a true negative occurrence than a true positive occurrence. The methods therefore for a two-class problem assign greater cost to false negatives than to false positives and thereby improving performance with regards to the positive class (Elkan, 2001; Nguyen et al., 2009). In cost-sensitive learning, cost-sensitive functions are either optimized directly or cost-insensitive algorithms converted to cost-sensitive algorithms by adopting various methodologies of weighting, thresholding, sampling, and ensemble learning (Alejo et al., 2007; Ling et al., 2004; Nguyen & Ho, 2005; Zhou & Liu, 2006). Drawbacks with cost-sensitive learning approach however include the assumption that the misclassification costs are known which is rarely the case in real situation. Cost-sensitive classifiers are also known to be prone to data over fitting during training (Weiss, 2004), and so extra care must be taken in the calibration stage with this approach.

2.6.1.5 Ensemble methods

Ensemble-based methods, also called multiple classifier systems (Polikar, 2006) are known to merge the performances of many classifiers to produce a single aggregate prediction which outperforms any other classifier considered individually (López et al., 2013; Phoungphol, 2013). Ensembles of classifiers have recently been presented as a viable solution to the imbalanced data distribution problem (Kuncheva & Rodríguez, 2014; Liu et al., 2009; Seiffert et al., 2010; Sun et al., 2007; Van Hulse et al., 2009; Wang & Yao, 2013). The ensemble frame work is usually built from combination of various existing ensemble learning algorithms and any of the earlier discussed approaches including mostly data resampling and cost sensitive learning. The commonly adopted ensemble learning algorithms are the bagging, boosting, voting and stacking algorithms (Jason, 2016; Kuncheva & Rodríguez, 2014), of which the bagging and boosting are the most popular.

Bagging, works by training individual classifier using different bootstrap of the data set (Breiman, 1996). The most widely known bagging algorithm is the random forest (Breiman, 2001). Boosting was proposed to train sequence of classifiers on difficult learning instances (Schapire, 1990, 2003). For an imbalanced data situation, boosting functions by iteratively uplifting classifier performance via updating misclassification cost or by modifying data distribution ratio (Chawla et al., 2003; Sun et al., 2007; Tang et al., 2009). A detailed classification of the ensemble methods for learning imbalanced data has been extensively described elsewhere (Galar et al., 2012). The study reported by (Galar et al., 2012) showed that classifiers ensemble-based results outperform results obtained from using data resampling techniques in conjunction with training a single classifier. Simple ensemble approaches like the RUSBoost and UnderBagging have also been reported to outperform many other more complex algorithms (Barandela et al., 2003; López et al., 2013; Seiffert et al., 2010).

2.7 Evaluation metrics for imbalanced data analysis

Evaluation metrics adopted are very critical for classification performance assessment and modelling guidance. Overall accuracy (well known traditionally) has been presented as inappropriate for measuring classifier performance in an imbalanced data situation (Liao, 2008; Nguyen et al., 2009), when considered alone. For example, a classifier might obtain 99% accuracy in an imbalanced data set comprising of 99% examples of the prevalent class. This kind of result is misleading and therefore other measures have been proposed for an imbalanced data distribution scenario. Such measures summarising the performance of a classifier are as shown in a confusion matrix displayed in Table 2.3. Other metrics of importance apart from those directly elucidated from the confusion matrix include: Precision, Recall, precision-recall curve, positive and negative

predictive value, F-measure, G-mean, Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC).

Table 2.3 Confusion matrix

		Prediction class	
		Predicted as Positive	**Predicted as Negative**
True class	**Actually Positive**	True Positives (TP)	False Negatives (FN)
	Actually Negative	False Positives (FP)	True Negatives (TN)

In a binary class classification situation, the particular class with very few training samples but with high identification importance is commonly referred to as the positive class and the other as the negative class (Sun et al., 2009). This definition however seems not to be directly applicable to most Agricultural and food processing operations. For example, even though non-fertile eggs in chicken egg fertility assessment study belongs to the rare class (very few training examples), fertile eggs of the majority class are of higher identification importance, from the hatchery industries point of view. Therefore, Agricultural and food processing applications might not fit in directly to the definition of positive class being the class with very few training samples and simultaneously of higher recognition importance. Nonetheless, the definitions of the acronyms in Table 2.3 remains unchanged with background understanding of class definitions. These definitions according to (François, 2006; Sun et al., 2009) are as described thus:

***True positive rate (TPR)*:** Proportion of actual positive instances that are predicted as positive

TPR = TP/ (TP+FN) * 100

***True negative rate (TNR)*:** Proportion of actual negative examples that are predicted as negative

TNR = TN/(TN+FP) * 100

***False positive rate (FPR)*:** Proportion of actual negative examples that are predicted as positive

FPR = FP/ (FP+TN) * 100

***False negative rate (FNR)*:** Proportion of actual positive instances that are predicted as negative

FNR = FN/ (FN+TP) * 100

Error rate (ERR) and Overall accuracy (OVA) can as well be computed from above as:

ERR = (FP+FN)/ (TP+FN+FP+TN) * 100

OVA = (TP+TN)/ (TP+FN+FP+TN) * 100 = 1 – ERR

***Sensitivity*:** This is also known as "recall" (R) in information retrieval systems or "true positive rate" (TPR) as earlier described

***Specificity*:** This is also known as "true negative rate" (TNR) as earlier described

***Positive Predictive Value (PPV)*:** Proportion of predicted positives that are actual positives. This is also called "precision" (P) in information retrieval systems. It must be noted that "precision" might also be described in terms of the negative predictive value in a situation where the rare class has not been taken as the positive class.

PPV = P = TP/ (TP+FP) * 100

***Negative Predictive Value (NPV)*:** Proportion of predicted negatives that are actual negatives

NPV = TN/ (TN+FN) * 100

***F-measure*:** When only the performance of the positive class is critical, two measures namely TPR or recall (R) and PPV or precision (P) are adequate. F-measure has been suggested by (Lewis & Gale, 1994) to integrate averagely these two measures. F-measure is therefore usually represented as the harmonic mean of precision and recall thus:

$$F - measure = \frac{2 * P * R}{\mathrm{P} + \mathrm{R}}$$

***G-mean*:** In situation where both performances of the positive and negative classes are paramount, both TPR and TNR are expected to be simultaneously high enough. Hence, (Kubat et al., 1998) proposed the G-mean metric to measure the balanced performance of classifier between two classes as:

$$G - mean = \sqrt{\mathrm{TPR} * \mathrm{TNR}} = \sqrt{\mathrm{Sensitivity} * \mathrm{Specificity}}$$

2.7.1 ROC analysis

ROC graphs have long been in existence and widely used in the field of signal theory and detection (Egan, 1975; Swets et al., 2000). It has been equally extended for use in visualizing and analysing behaviour of diagnostic systems (Swets, 1988). The earliest use in machine learning was however traced to the work of (Spackman, 1989) who evidently revealed the potential of ROC curves in evaluating and comparing algorithms. The machine learning community in recent times have witnessed an increase in the use of ROC charts partly due to the understanding that the conventional overall accuracy approach is a substandard yardstick for performance evaluation (Provost & Fawcett, 1997; Provost et al., 1998). ROC graphs have been shown to be specifically useful in the skewed class distribution and unequal classification error costs domains. These

attributes have made ROC analysis increasingly important especially in the present emerging fields of cost sensitive and imbalanced data learning (Fawcett, 2006).

ROC analysis examines the interrelationship between sensitivity (TPR) and specificity (TNR) of a binary classifier. Due to prediction changes from score threshold variation, measurements in pairs (FPR, TPR) are generated for each selected singular threshold value (Sun et al., 2009). These measurements are connected in a Receiver Operating Characteristic (ROC) curves, having the true positive rate (TPR) on the Y-axis and the false positive rate (FPR), usually denoted as one minus true negative rate (1-specificity), on the X-axis (Figure 2.13). The optimal

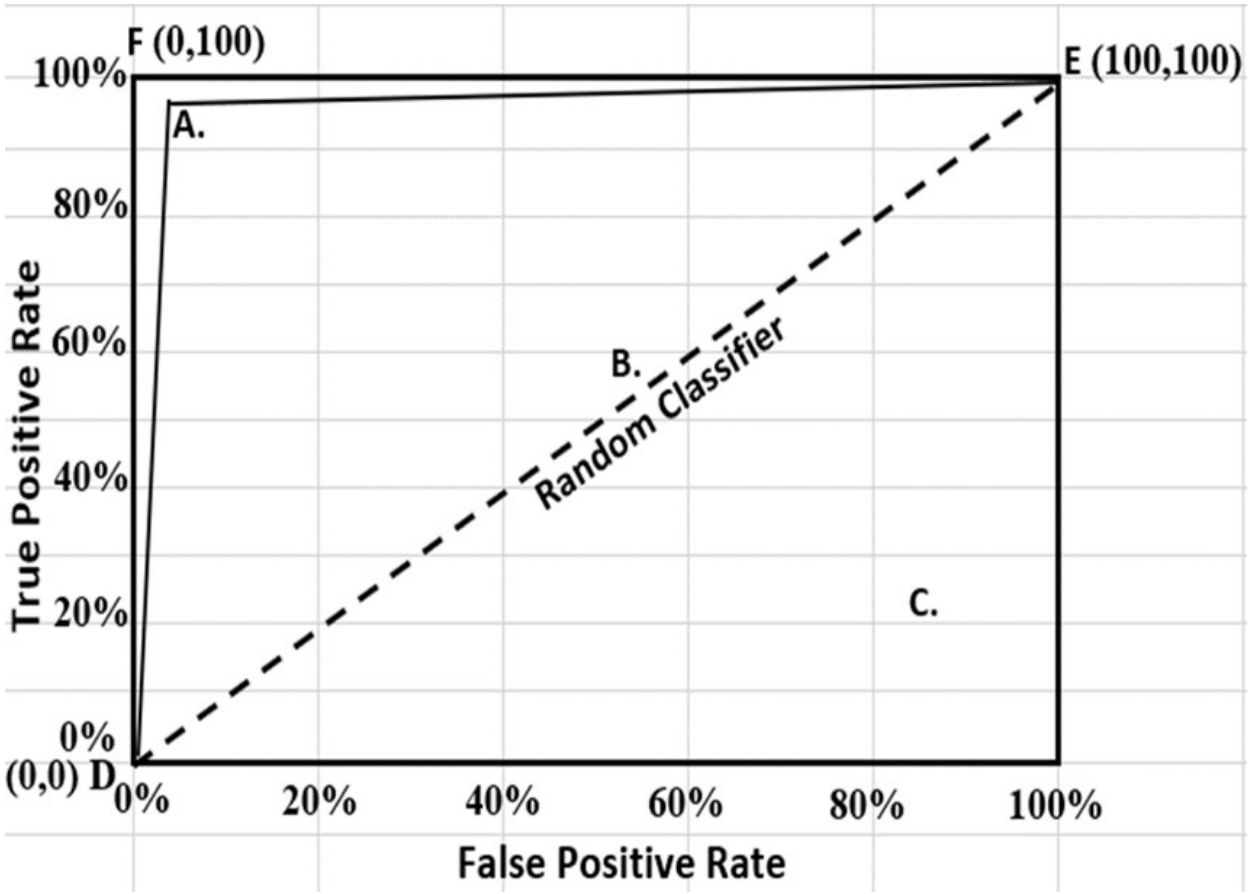


Fig. 2.13 Receiver Operating Characteristic (ROC) curves for different classifiers Modified from (López et al., 2013).

Classifier “F”, the ideal or perfect model is that which achieves false positive rate of 0% but sensitivity or true positive rate of 100% (FPR = 0, TPR = 100). Hence, a good classification model is usually positioned as close as possible to the upper left corner of the graph such as model “A”, while a model making a random guess would be located along the main diagonal (DBE), connecting the points (TPR = 0, FPR = 0) and (TPR = 100, FPR = 100). Therefore, any model positioned on the diagonal such as model “B” or below the diagonal like model “C”, are considered poor. ROC is thereby shown to depict relative trade-offs between costs (false positives) and gains (true positives). Further description of ROC curves and its implementation can be obtained from (Altman & Bland, 1994; Brown & Davis, 2006; Fawcett, 2006; López et al., 2013).

2.7.2 Area under ROC curve (AUC)

Since ROC curves show two-dimensional representation of classifier performance, there is usually a need during classifiers comparison analysis, to reduce ROC performance to a single scalar value depicting the expected performance (Fawcett, 2006). AUC gives such singular measure of a classifier’s performance for investigating which model is preferable on the average (Bradley, 1997; Hanley & McNeil, 1982; López et al., 2013). AUC being a portion of the area of the 100%-unit square (Fig. 2.13), will always have values between 0 and 100%. However, having the random guessing positions on the diagonal line between points (0,0) and (100,100) with an area of 50% or 0.5, there cannot be any good classifier with an AUC that is less than 50% (Fawcett, 2006). Xia et al. (2013) in a metabolomic biomarker discovery study, assessed utility of model features based on AUC values (%) as follows: 90-100 = excellent; 80-90 = good; 70-80 = fair; 60-70 = poor; and 50-60 = fail.

2.7.3 Precision-recall curve

There exists situation where there is a need for both precision and recall being high enough and hence necessitating the determination of a safe threshold value for this determination. The trade- off between precision and recall in such situation can be easily observed using the precision-recall curve as depicted in Figure 2.14 (Davis & Goadrich, 2006; Raghavan et al., 1989; Saito & Rehmsmeier, 2015).

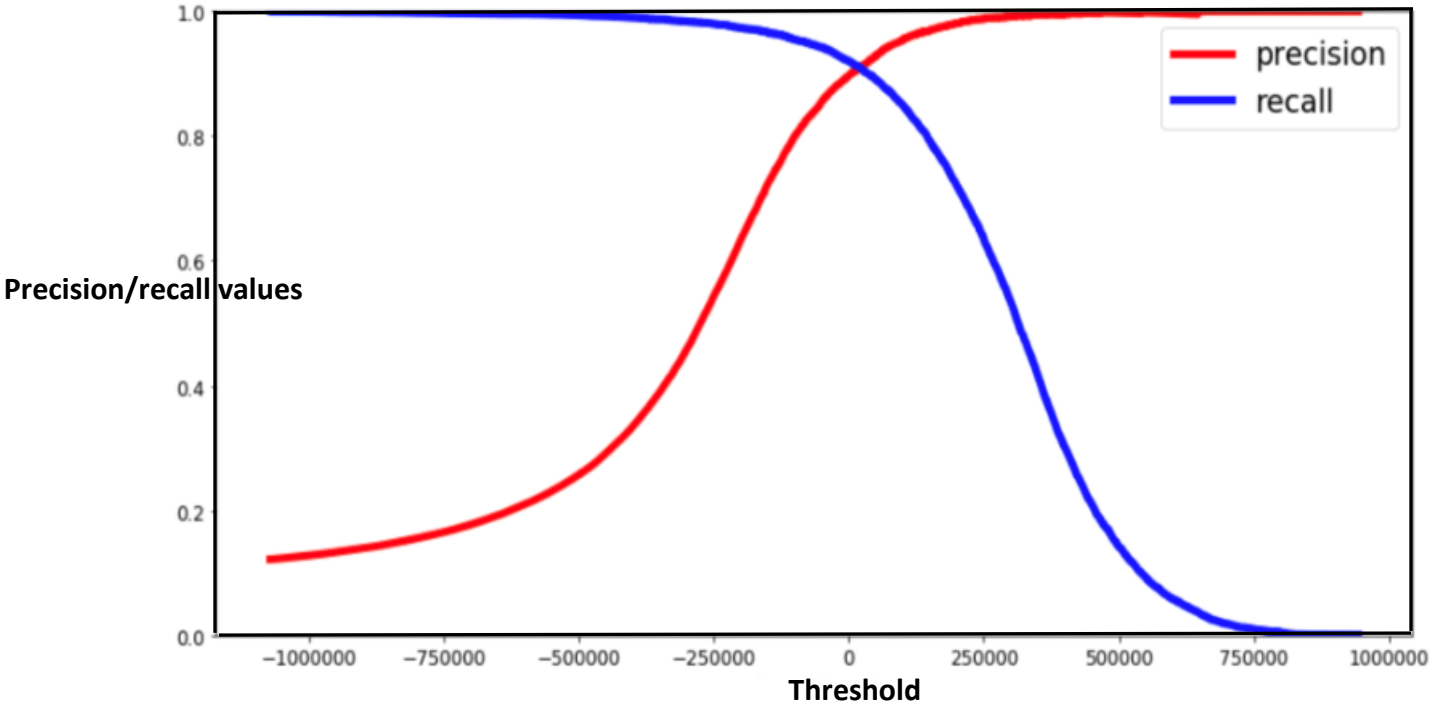


Fig. 2.14 Typical precision-recall curve for best threshold identification

Source: (https://machinelearning-blog.com/2018/04/03/evaluation-metrics-for-classification/)

2.8 Conclusion

This review has focused on the present direction in the use of non-destructive hyperspectral imaging technology for food quality analysis, towards improvement in the use of non-destructive technologies for chicken egg fertility assessment. The chapter commenced by reviewing the

present state of chicken egg production, moved on to discussion on egg formation and structure with exposition on chicken egg chemical and functional compositions. The intrinsic nature of chicken egg fertility and the industrial challenge of identifying fertility prior to incubation were thoroughly examined. Existing methods of assessing chicken egg fertility were then investigated with a view to proffering state-of-the-art solution(s) to the identified problems existing in chicken egg fertility data structure during multivariate analysis. Three major areas needing attention towards building a futuristic robust model for chicken egg fertility early detection were eventually enunciated as sample size, analysis/modelling techniques, and the rare class data acquisition problem. Feature extraction/selection techniques and addressing imbalanced data problem via various resampling, one-class learning, cost sensitive learning, and ensemble approaches were proposed for uplifting existing results with chicken egg fertility early discrimination. It is believed that the appropriate implementation of the outcome of this review would assist tremendously the commercial hatchery industries towards achieving a stable and robust model for chicken egg fertility early discrimination. Furthermore, the outcome of this review would find useful applications in appropriate handling of Agro-food imbalanced data during multivariate down stream analysis.

CONNECTING TEXT TO CHAPTER 3

Chapter two of this thesis reviewed the current direction in the use of non-destructive hyperspectral imaging technology for food quality analysis in general and specifically, chicken egg fertility assessment, using chemometric and machine learning approaches. The literatures search showed there was no existing robust model for early chicken egg fertility classification especially prior to incubation. The need to enhance existing modelling, validation and verification approaches was identified and examined in the review. Chapter 3 therefore sought to build a robust model for early chicken egg discrimination, using hyperspectral imaging, in conjunction with a supervised learning partial least square (PLS) regression approach. The effect of varying threshold values on classification accuracies was studied on naturally imbalanced brown and white chicken egg data.

Chapter 3 would be submitted for publication to the Journal of Chemometrics and Intelligent Laboratory Systems as:

Adegbenjo, A. O., Liu, L., and Ngadi, M. (2019). An adaptive partial least square (PLS) regression approach for classifying chicken egg fertility hyperspectral imaging data. *Chemometrics and Intelligent Laboratory Systems*

CHAPTER 3

A PLS REGRESSION TECHNIQUE FOR CLASSIFYING CHICKEN EGG FERTILITY HYPERSPECTRAL IMAGING DATA

Abstract

Partial least square (PLS) regression is a well-known chemometric method used for predictive modelling, especially in the presence of many variables. Although PLS was not initially developed as a technique for classification tasks; scientists have reportedly used this approach successfully for discrimination purposes. Whereas some non-supervised learning approaches including but not limited to PCA, and k-means clustering do well in identifying/understanding grouping and clustering patterns in multidimensional data, they are limited when the end target is discrimination, making PLS a preferable alternative. A total of fertilized 672 chicken egg hyperspectral imaging data, consisting of 336 white eggs and 336 brown eggs were used in this study. Hyperspectral images in the NIR region of 900-1700 nm wavelength range were captured prior to incubation on day 0 and on days 1-4 after incubation. Eggs were candled on incubation day 5 and broken out on day 10 to confirm fertility. While a total number of 312 and 314 eggs were found to be fertile in the brown and white egg batches respectively, total numbers of non-fertile eggs in the same set of batches were 23 and 21 respectively. Spectral information was extracted from a segmented region of interest (ROI) of each hyperspectral image and spectral transmission characteristics were obtained by averaging the spectral information. Threshold values were varied between 0.50-0.85 for discrimination by implementing the PLS regression algorithm on the calibration set, at each selected threshold value. With true positive rates (TPR) of up to 100% obtained at considered threshold values and on different days of incubation, the results indicated that the proposed PLS

technique can discriminate between fertile and non-fertile eggs. The adaptive PLS approach was thereby presented as suitable for handling hyperspectral imaging-based chicken egg fertility data.

3.1 Introduction

Out of 13.1 billion hatching eggs produced in the U.S. egg industry for the year 2005, the ratio of layer to broiler eggs produced was reported to be around 12:1, creating different degrees of discriminating tasks to both layers and broilers industries. With fertility rates in the range of 60 to 90% (Lawrence et al., 2006; NASS, 2006), there could be about 1.3 billion to over 5 billion infertile eggs being incubated yearly. According to the Agriculture and Agri-Food Canada report 2013, total hatching egg product was set at 798.3 million, resulting in a minimum of about 80 million non-fertile eggs being incubated annually in Canada alone which is worth a whopping sum of about $27.6 million being lost annually. Furthermore, discarding of non-hatching eggs has consistently posed significant disposal problems for the hatcheries, especially in the case of exploder eggs in hatching cabinet, resulting in high tendency of molds and bacteria infestation to other eggs (Lawrence et al., 2006). Thus, identification and isolation of infertile eggs from fertile eggs have significant economic and safety implications for commercial broiler breeders.

Recent researches indeed have supported the great potential applications of Hyperspectral imaging as a non-destructive method for assessing fertility/hatchability, embryo development and mortality rates in chicken eggs. These studies have however reported mostly on fertility detection of white-shelled chicken eggs with scanty reports on brown eggs and where available, results were not as promising as with white eggs. Also, samples considered in earlier studies were small (Lawrence et al., 2006; Liu & Ngadi, 2013; Smith et al., 2008). Smith et al. (2008) reported low validation and verification accuracies for fertility detection in brown eggs (Validation data sets: 71% for Day 0; 63% for Day 1, 65% for Day 2, 83% for Day 3; Verification data sets: 51% for

Day 0 and 50% for Day 3). It was concluded that the Mahalanobis Distance (MD)/Principal Component Analysis (PCA) model used was not adequate for the classification. This is of a great concern for the poultry industry as this means large number of fertile eggs would end up being discarded based on such model. Hence, there is indeed an urgent need for more appropriate classification technique for egg fertility assessment.

Partial least square (PLS) regression, also commonly known as the Projection to Latent Structure (Swarbrick, 2012) is a widely used technique that have found useful applications in various domains including but not limited to the engineering, medicine, and agriculture. The PLS approach is particularly known for building predictive models with many variables rather than explaining underlying correlations between variables (Yu, 2000). PLS was not initially developed as an approach for statistical classification tasks except for regression; nonetheless, scientists have reportedly used this approach successfully for discrimination purposes (Barker & Rayens, 2003; Briandet et al., 1996; Gottfries et al., 1995; Iizuka & Aishima, 1997; Ortiz et al., 1996). Even though principal component analysis (PCA) is a well-known chemometric method that has recorded notable success as a pre-classification procedure, this success has been reported to be only possible in various application domains because of its favourable disposition to consider the among-groups variability rather than the within-groups variability (Barker & Rayens, 2003). This mode of PCA implementation therefore do not address situation in which the within-groups variability in data, is also of a major concern due to existence of several sub-clusters in a single class, not having the same number of samples (Japkowicz, 2001). The within-class variability occurrence has been reported to have unfavourable consequence on learning algorithms (Yoon & Kwek, 2007). In such situation, PCA has been observed to perform less optimally and thereby presenting PLS as the next applicable alternative. According to (Barker & Rayens, 2003), PLS was

reported to have potential of outperforming PCA when within-groups variability dominates the among-groups variability. Additionally, PLS has been judged versatile in solving data structural problems like skew distributions, multicollinearity, and missing regressors condition- all which are peculiar characteristics of hyperspectral imaging data (Cassel et al., 1999).

(Liu & Ngadi, 2013) reported a perfect classification accuracy using PCA and k-means clustering. However, these approaches being non-supervised learning techniques need further confirmation using standard supervised learning algorithm(s). Although some non-supervised learning approaches like the PCA, k-means, k-medians, and hierarchical cluster analysis are effective with identifying/understanding grouping and clustering patterns in multidimensional data, they are limited when the end target is discrimination (Barker & Rayens, 2003). Furthermore, unsupervised classification is always the starting point in any discrimination problem and should necessarily be followed by supervised classification (Swarbrick, 2012), towards an industrial adoptability consideration.

In view of the foregoing, this study has therefore examined, the suitability of an adaptive supervised learning PLS regression approach, together with a threshold-moving technique, in handling chicken egg fertility hyperspectral imaging data. Classification accuracy have been based on the confusion matrix evaluation criterion at the expense of the more general overall accuracy computation, which has been shown to be inappropriate when dealing with data containing a rare class (Liao, 2008) as with the non-fertile eggs in chicken egg fertility data.

3.2 Materials and Methods

3.2.1 Samples

A total of 336 Brown shell eggs and 336 White shell eggs were received from a commercial fertile egg producer (Simetin Hatchery; www.couvoir.com) in 14 batches (48 eggs per batch) over

a period of 3 months. There were 7 batches of eggs collected in each group of brown and white egg sets. Tables 3.1 and 3.2 show the details of the overall egg samples available for analysis on each day of incubation for both brown and white eggs respectively. Out of the total 336 eggs received for both brown and white eggs, the number of total available eggs eventually used for analysis varied with incubation time due to egg breakage during handling. While 2 eggs (1, day 0; 1, day1) were broken from the brown egg batch, a total of 3 eggs (1, day 0; 2, day 3) were broken from the white egg batch. This variation in total available eggs during analysis results into a slightly different degree of imbalance from one day of incubation to another. The ratio of non-fertile to fertile eggs in this study is estimated from Tables 3.1 and 3.2 to be 1:13 and 1:15 for both brown and white eggs respectively.

3.2.2 Image acquisition and processing

A laboratory near-infrared (NIR) hyperspectral imaging system used in this project comprised of an InGaAs camera, a conveyor (Donner 2200 series, Donner Mfg. Corp., USA) driven by a stepping motor (MDIP22314, Intelligent motion system Inc., USA), a line-scan spectrograph *(*HyperspecTM, Headwall Photonics Inc. USA) with a NIR spectral wavelength range from 900 to 1700 nm and a spectral resolution of 4.79 nm, a tungsten halogen lamp (50 W) providing back illumination to eggs, an enclosure supporting the system, a data acquisition and pre-processing software *(*Hyperspec, Headwall Photonics Inc. USA) and a PC. All eggs were first imaged by the hyperspectral imaging system on Day 0 (just prior to incubation) and immediately after imaging, the eggs were incubated in an Ova-Easy 190 Advance Series II Cabinet Incubator (Brinsea Products Inc., Florida, USA) at 37.78°C (100°F) and 55% relative humidity. The eggs were automatically turned every hour. On days 1, 2, 3, and 4 of incubation, eggs were removed for imaging in sequence and then immediately returned into the incubator, in a process of about 1 min.

Table 3.1 Overall egg sample specifications for brown eggs

Incubation period	Egg received	Broken	Total Eggs used	Fertile (F)	Non-fertile (NF)
Day 0	336	1	335	312 (93.13%)	23 (6.87%)
Day 1	335	1	334	311 (93.11%)	23 (6.89%)
Day 2	334	-	334	311 (93.11%)	23 (6.89%)
Day 3	334	-	334	311 (93.11%)	23 (6.89%)
Day 4	334	-	334	311 (93.11%)	23 (6.89%)

Table 3.2 Overall egg sample specifications for white eggs

Incubation period	Egg received	Broken	Total Eggs used	Fertile (F)	Non-fertile (NF)
Day 0	336	1	335	314 (93.73%)	21 (6.27%)
Day 1	335	-	335	314 (93.73%)	21 (6.27%)
Day 2	335	-	335	314 (93.73%)	21 (6.27%)
Day 3	335	2	333	312 (93.69%)	21 (6.31%)
Day 4	333	-	333	312 (93.69%)	21 (6.31%)

After 10 days of incubation, eggs were broken out to determine fertility. The output hypercube image obtained is 800 rows x 320 columns x 167 bands. The region of interest (ROI) of obtained spectral images was individual egg of each sample image. The ROI was selected at maximum wavelength band 37 (1071 nm) and punched through other wave bands. A mask was created for each individual egg to segment it from the original spectral image that normally included four eggs. Segmented individual eggs were then used for calculating mean spectra, following standard procedures.

3.2.3 Spectral transmission and feature extraction

Spectral transmission characteristic namely Mean Spectral, MS and extracted features based on thresholding were used in this study for further data analysis. MS stands for the mean

value of all pixels in ROI for the current wavelength over the spectral range of 900-1700 nm. Threshold-moving method has been used in cost-sensitive neural networks learning with reported good effectiveness even with highly imbalanced data sets. More detailed explanation of this method is as described by (Longadge & Dongre, 2013; Williams et al., 2009). Threshold (TR) values considered for extraction of features in the present work ranged between 0.50 - 0.85. The purpose of adopting thresholding technique in conjunction with PLS algorithm was to extract useful spectral features to facilitate the discrimination of fertile eggs from non-fertile eggs. With the choice of an appropriate threshold value, a new set of features with the potential of achieving optimal classification accuracy is extracted for analysis, and discrimination performance was then evaluated using the confusion matrix evaluation criterion. All operations were performed in the MATLAB R2014a (The MathWorks, Inc., MA, USA) platform.

3.2.4 Partial least square regression analysis

For the different days of egg incubation, a PLS code written in the MATLAB R2014a environment was used for data analysis and full cross validation was later employed as a means of internal validation in all cases. Unlike the popular multiple linear regression (MLR) which is prone to the problem of over-fitting in the presence of too many factors, PLS analysis do adjust to over-fitting problem by extracting only the latent factors accounting for majority of the manifest factor variation (Tobias, 1995). Not only this, PLS is well known for analyzing data with strongly collinear (correlated), noisy, and numerous X-variables. The PLS analysis as adopted in this study models both the X- and Y-matrices simultaneously (thereby maximizing the covariance between X and Y) to reveal the latent variables in X, having the potential of predicting accurately the latent variables in Y (Wold et al., 2001). Unlike the PCA, which decomposes X to obtain components

that explains most variability in X, PLS seeks to identify components from X that best predict Y (Abdi, 2010).

3.2.5 Choice of optimal number of PLS components (PCs)

PLS modelling process is greatly influenced by only few underlying (latent) variables; whereas, the appropriate number of these latent variables is usually unknown. One major aim of PLS analysis therefore was to estimate this number (Wold et al., 2001) and in doing so, it becomes very critical to identify an optimum value for the user- defined number "n" of PLS components (PCs). which is directly related to the selection of informative features required for accurate discrimination process. This study have followed the full (leave one out) cross-validation (CV) procedure reported by (Wold et al., 2001) in testing the predictive performance of PLS components and stopping when adding more components tends to reduce performance. Detailed explanation of this procedure has been described elsewhere (Clark & Cramer, 1993; Höskuldsson, 1988, 1996; Wakeling & Morris, 1993; Wold et al., 1993). Nevertheless, because the end goal in this study is discrimination and not regression, the traditional interpretation of the CV procedure cannot be applied directly in entirety and the reason why the confusion matrix criterion was adopted for evaluating discrimination performance. PCs ranging from n = 5 to n = 50 (in interval of 5) were tested for classification accuracy before arriving at an optimum value for "n". The threshold for initial feature extraction was chosen to be 0.80 from preliminary trial and error analysis. This threshold was subsequently used in predictive performance testing for determining optimum number of PCs.

3.2.6 Criteria for evaluating discrimination performance

Overall accuracy has been presented as inappropriate for measuring classifier performance in the situation consisting of a rare class data (Liao, 2008; Nguyen et al., 2009). The present study

has therefore adopted the confusion matrix evaluation criterion for a binary-class egg fertility discrimination problem**.** In a binary class classification situation, the particular class with very few training samples but with high identification importance is commonly referred to as the positive class and the other as the negative class (Sun et al., 2009). This definition however seems not to be directly applicable to most Agricultural and food processing operations. Even though non-fertile eggs in this research belongs to the rare class (very few training examples), fertile eggs of the majority class are of higher identification importance, from the hatchery industries point of view. Therefore, Agricultural and food processing applications might not fit in directly to the definition of positive class being the class with very few training samples and simultaneously of higher recognition importance. Nonetheless, in this first study, we have maintained taking non-fertile eggs as the positive class not only because they fall into the minority class, but also that the future industrial instrumentation for egg fertility assessment might be much more economically built and viable to reject non-fertile eggs (fewer samples) than accepting fertile eggs (larger number of samples). The choice of our true positive class in this first study is critical to be able to examine our results while maintaining conventional consistency.

The confusion matrix employed for the interpretation of the PLS analysis and hence determining classification accuracy is as shown in Table 3.3. If TP = True positive (number of non-fertile eggs classified as non-fertile), TN= True negative (number of fertile eggs classified as fertile), FP= False positive (number of fertile eggs classified as non-fertile), and FN=False negative (number of non-fertile eggs classified as fertile), the following equations 3.1-3.4 as reported in (François, 2006; Sokolova & Lapalme, 2009; Sun et al., 2009) can be obtained, where TPR, FPR, TNR, and FNR represent the rate in percentage of true positive examples, the rate in percentage of false positive examples, the rate in percentage of true negative examples, and the rate in percentage of

false negative examples, respectively. The traditional overall accuracy (OVA), including the error rate (ERR) can also be computed as shown in equations 3.5 and 3.6.

Table 3.3 Confusion matrix

		Prediction class	
		Predicted as Positive	**Predicted as Negative**
True class	**Actually Positive**	TP	FN
	Actually Negative	FP	TN

$$TPR = TP/ (TP+FN) * 100 \quad (3.1)$$

$$TNR = TN/(TN+FP) * 100 \quad (3.2)$$

$$FPR = FP/ (FP+TN) * 100 \quad (3.3)$$

$$FNR = FN/ (FN+TP) * 100 \quad (3.4)$$

$$OVA = (TP+TN)/ (TP+FN+FP+TN) * 100 \quad (3.5)$$

$$ERR = (FP+FN)/ (TP+FN+FP+TN) * 100 \quad (3.6)$$

3.3 Results and Discussion

While a total number of 312 and 314 eggs were found to be fertile (F) in the brown and white egg batches respectively, total numbers of non-fertile (NF) eggs in the same set of batches were 23 and 21 respectively (see Tables 3.1 and 3.2), at the start of our analysis. Figure 3.1 showed typical transmittance MS profiles of brown eggs, on different days of incubation. It was observed from Figure 3.1 that fertile eggs maximum transmittance intensity decreases as incubation period increases from day 0 through day 3. This observation seems related to the onset of active molecular

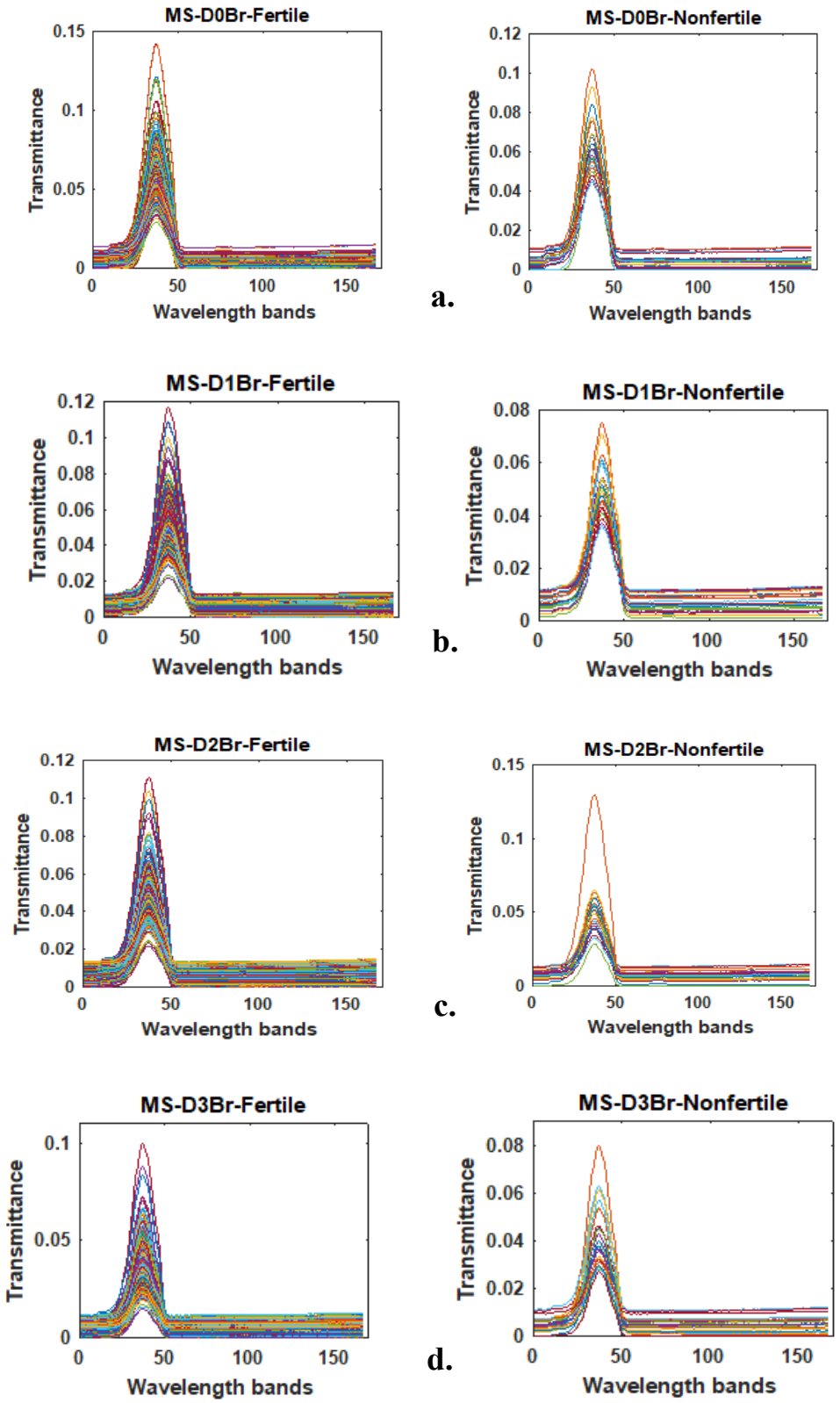
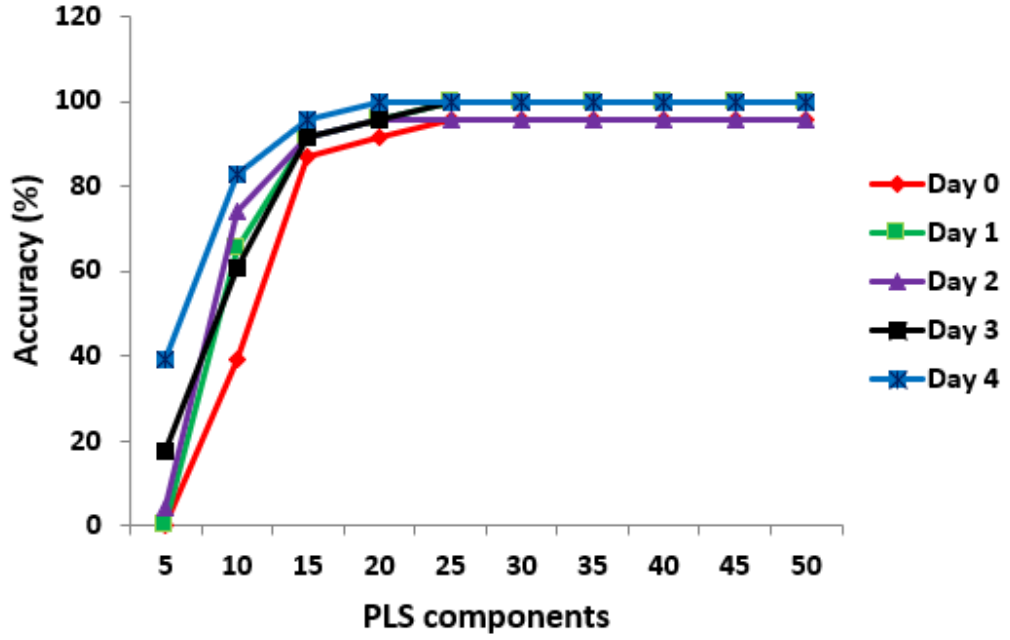


Fig. 3.1 Typical transmittance mean spectra (MS) profiles of brown eggs, on different days of incubation (a), prior incubation (b), day 1 incubation (c), day 2 incubation (d) day 3 incubation

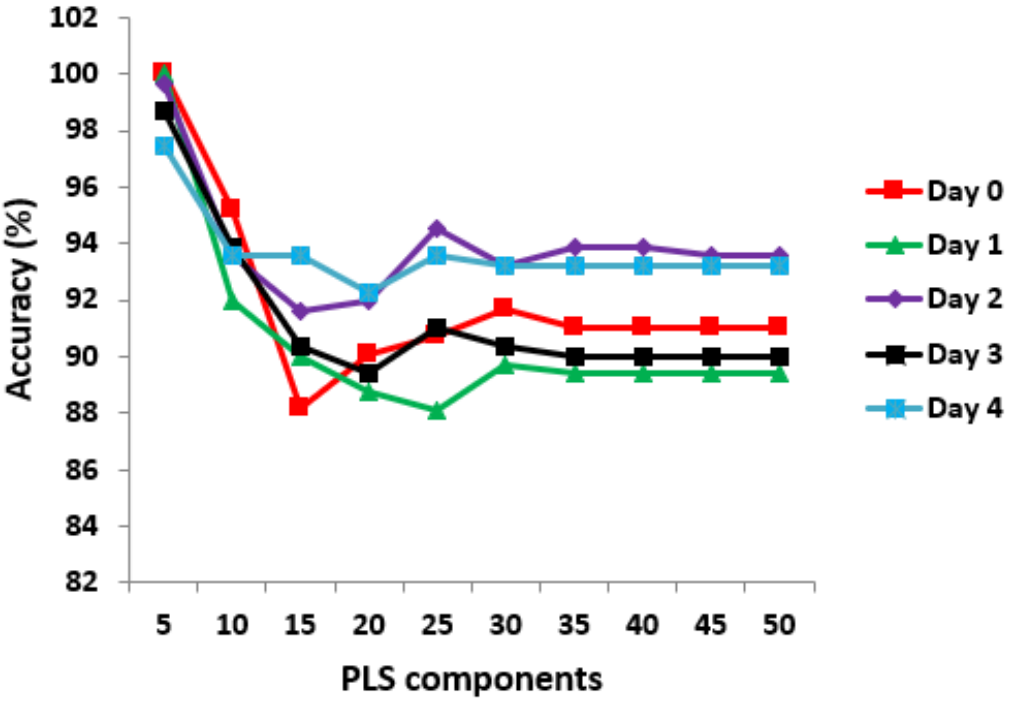
activities from meiotic and mitotic cell divisions in the fertile eggs. Knowing that the proportion of light absorbed by any material is dependent on the quantity of molecules involved in molecular interaction, fertile eggs tend to absorb more light at different wavelengths as incubation period progresses, and hence the amount of light being transmitted decreases accordingly. Non-fertile eggs transmission intensity over the considered incubation periods did not follow a definite trend. The initial decrease in maximum transmittance intensity from day 0 to day 1 might as well be related to the molecular interactions from meiotic cellular division. As meiosis process terminated in the non-fertile eggs, further cell division also ceased, since there was no fertilization to trigger the onset of mitotic cell division. Therefore, subsequent egg maximum intensity increase, and later decrease can be attributed to the degree of albumen-yolk solution concentration, and this is dependent on the rate of yolk dissolution into the albumen under incubation conditions. From Beer's law, solution concentration is directly proportional to light's absorption (Norris, 1996; Williams & Norris, 1987) up to a specific level, as the law failed at some higher concentrations.

Using brown eggs data, Figure 3.2 shows the predictive performance chart for determining optimum number of PLS components for different days of egg incubation. The TPR performance chart (Figure 3.2a) showed that adding more components above 25 does not bring any further improvement in classification accuracy. 25 PLS components were then chosen for feature extraction and subsequent discrimination based on the TPR performance results. However, if TNR is of greater or equal interest, only the first 5 PLS components will suffice for further feature extraction (see Figure 3.2b). In the light of the above, further analysis in this study have used both 25 and 5 PLS components at various selected thresholds between 0.50 to 0.80 for feature extraction and eventual classification. Figure 3.3 showed evaluation metrics at threshold point 0.80 for both brown and white eggs, on different days of incubation, and with associated misclassification error

rates.



a.



b.

Fig. 3.2 Determining the optimum number of PLS components for brown eggs based on (a), TPR and (b), TNR

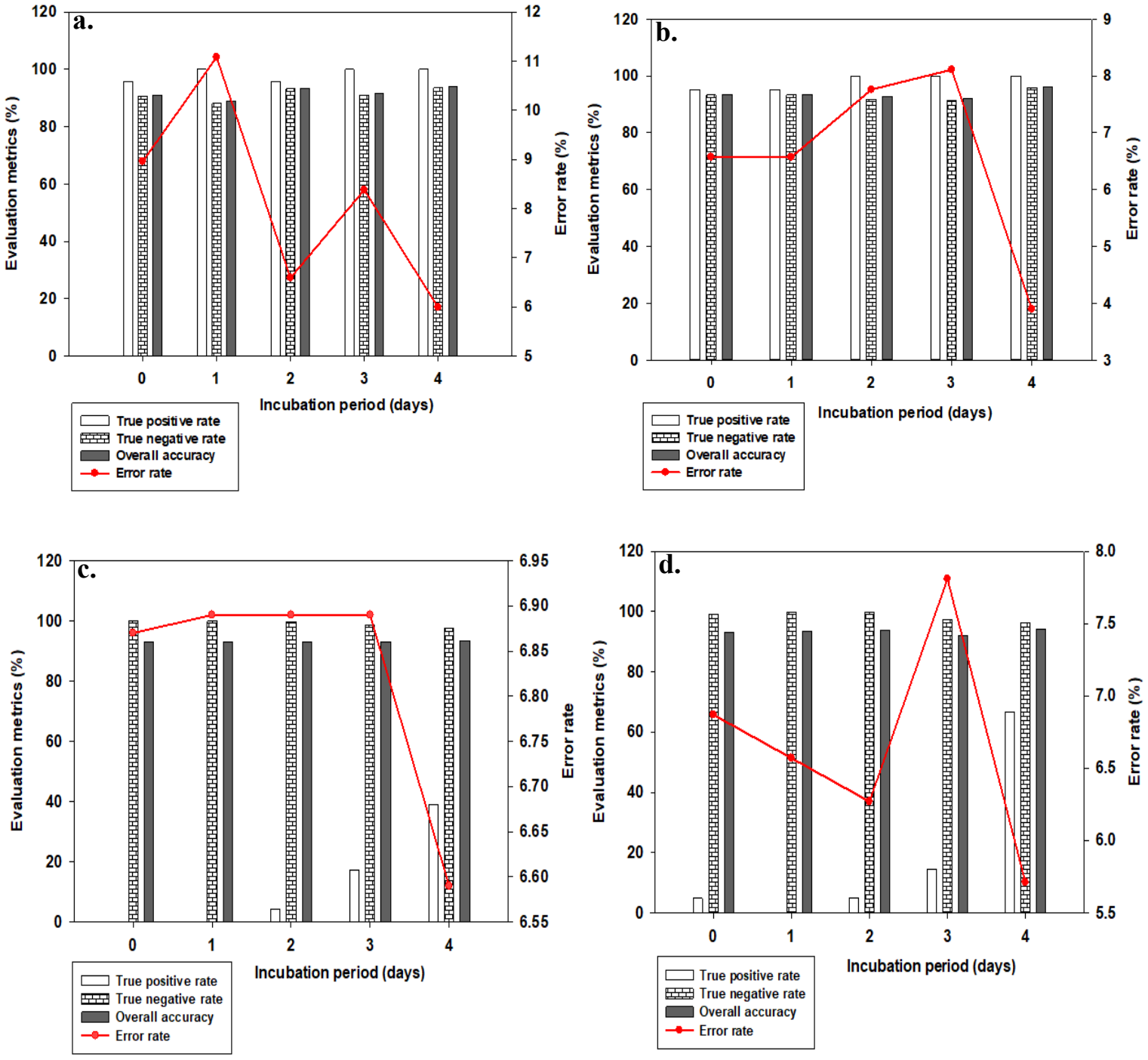


Fig. 3.3 Evaluation metrics (%) for built models on different days of incubation (a) brown eggs, 25 PCs (b) white eggs, 25 PCs (c) brown eggs, 5 PCs (d) white eggs, 5 PCs

From Figures 3.3a and 3.3b at 25 PCs, both brown and white eggs achieved 100% TPR accuracy on days 3 and 4. While none of the two sets of eggs achieved 100% TPR accuracy on day 0 incubation, 100% TPR accuracy was also obtained for both brown and white eggs respectively on incubation days 1 and 2. At 5 PCs (Figures 3.3c and 3.3d), none of the two sets of eggs achieved perfect TPR accuracy of 100% on all incubation days considered. 100% TNR accuracy was however obtained on days 0 and 1 for brown eggs, but also at a detrimental 0% TPR corresponding accuracies. Hence, the classifier despite classifying all fertile eggs as fertile, will also end up misclassifying all non-fertile eggs as fertile on these two days of incubation. The least misclassification error rates for the classes of eggs at both number of PCs considered were on day 4 incubation, but this day is already becoming too late for early recognition and classification, leaving us to consider earlier days (especially day 0) more critically. We indeed need a classifier mode that will perform at much closer margin of TPR and TNR accuracies and at the same time using much lesser number of PCs.

Figure 3.4 showed model accuracies and error rates on day 0 incubation, for all PCs considered from 5 to 50. It was observed that using PCs above 5 poses risk to model's robustness, as the misclassification error rates are found to increase after 5PCs.

Table 3.4 showed typical day 0 incubation confusion matrix results at some other selected thresholds 0.81 and 0.55, for both brown and white eggs. Detailed values used for the computation of the confusion matrices is as shown in appendix A. On day 0 of incubation for brown eggs (Table 3.4c), TPR of 100% was achieved at threshold value of 0.81; whereas white eggs achieved TPR classification accuracy of 95.24% at this same threshold (Table 3.4d). Detailed percent classification accuracy information for both brown and white eggs are as shown in appendices A1-A4, for both 25 and 5 PCs respectively. The same TPR classification accuracy of 95.24% was also

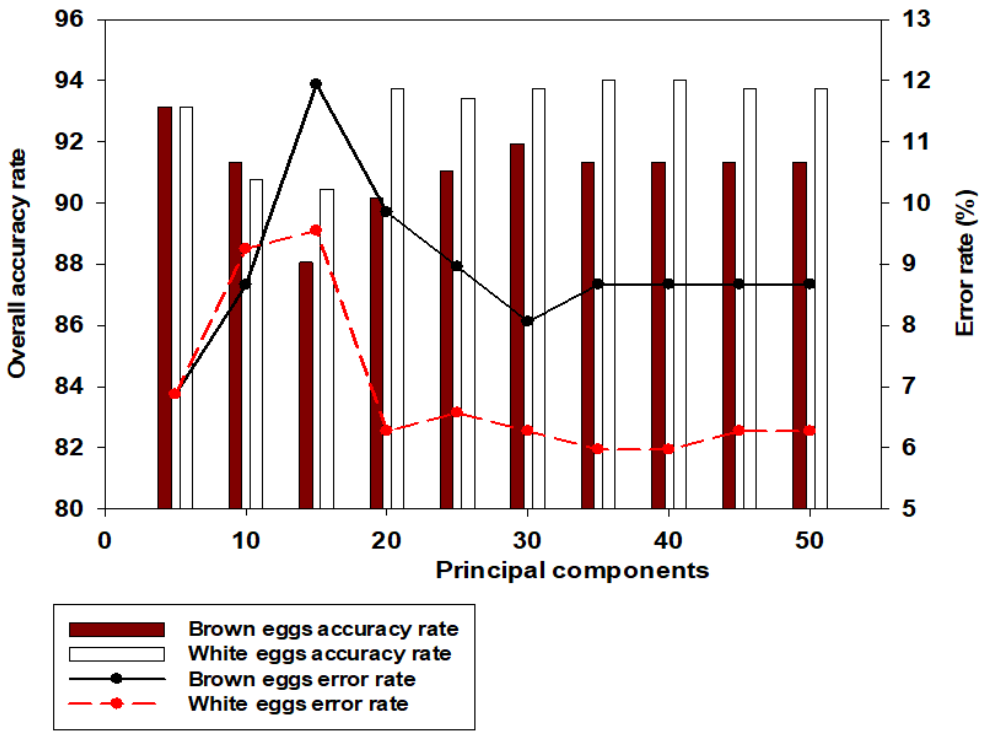


Fig. 3.4 Model accuracies and error rates for all PCs from 5 to 50 on day 0 incubation

achieved at TR 0.80 for white eggs, in which only one non-fertile egg was misclassified as fertile (see appendix A2). None of the threshold values considered between 0.50 to 0.85 for the white eggs, on day 0 incubation achieved 100% classification accuracy considering the TPR values. Considering the true negative rates (TNR) however, white eggs achieved accuracy of 100% at four threshold values of 0.5, 0.55, 0.60 and 0.65; whereas brown eggs achieved 99.68, 99.68, 99.68 and 99.04%, respectively at these same thresholds (Appendices A1, A2, and Tables 3.4a, b). For the brown eggs, only one fertile egg was misclassified as non-fertile at thresholds 0.5, 0.55 and 0.60, but three fertile eggs were misclassified as non-fertile at threshold value of 0.65. These results showed that the PLS algorithm used can discriminate both brown and white fertile eggs from non-fertile eggs prior to incubation using any of the thresholds identified.

Table 3.4 Typical day 0 confusion matrix for selected egg models at different thresholds and PCs (a) brown, TR 0.55, PC 25 (b) white, TR 0.55, PC 25 (c) brown, TR 0.81, PC 25 (d) white, TR 0.81, PC 25 (e) brown, TR 0.55, PC 5 (f) white, TR 0.55, PC 5 (g) brown, TR 0.81, PC 5 (h) white, TR 0.81, PC 5

(a)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	86.96 (20/23)	13.04 (3/23)
Actually Negative	0.32 (1/312)	99.68 (311/312)

OVA = 98.81%

(b)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	61.90 (13/21)	38.10 (8/21)
Actually Negative	0.00 (0/314)	100.00 (314/314)

OVA = 97.61%

(c)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	100.00 (23/23)	0.00 (0/23)
Actually Negative	9.62 (30/312)	90.38 (282/312)

OVA = 91.04%

(d)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	95.24 (20/21)	4.76 (1/21)
Actually Negative	7.32 (23/314)	92.68 (291/314)

OVA = 92.84%

(e)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	0.00 (0/23)	100.00 (23/23)
Actually Negative	0.00 (0/312)	100.00 (312/312)

OVA = 93.13%

(f)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	0.00 (0/21)	100.00 (21/21)
Actually Negative	0.00 (0/314)	100.00 (314/314)

OVA = 93.73%

(g)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	0.00 (0/23)	100.00 (23/23)
Actually Negative	0.00 (0/312)	100.00 (312/312)

OVA = 93.13%

(h)

True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	4.76 (1/21)	95.24 (20/21)
Actually Negative	0.00 (0/314)	98.73 (310/314)

OVA = 92.84%

The results shown specifically in Tables 3.4a-d at 25 PCs are promising for both brown and white eggs, considering closer margin of TPR and TNR accuracies. However, models built with 5 PCs as shown in Tables 3.4e-h are much in favour of the prevalent class as can be seen in the TNR perfect accuracies as against the TPR lowest accuracies. This observation has been reported in literatures to be related to imbalanced data phenomenon (He & Garcia, 2009; Mani & Zhang, 2003; Sun et al., 2009) in the considered data sets. Therefore, despite the overall percentage accuracy (OVA) obtained for brown eggs on day 0 incubation at various thresholds from 0.50

through 0.75 at 25 PCs were much higher than that at thresholds 0.80 and 0.81 (see Appendix A1), the final accepted model might not be based on this overall accuracy due to the imbalanced data phenomenon, shifting the overall accuracy performance in favour of the majority class at the expense of the minority class. This is the reason why performance is better judged based on the true positive and/or true negative rates. In the specific situation under discussion, it might be more appropriate to adopt a model based on the 0.81 TR (TPR value of 100%, OVA of 91.04%), than a model based on 0.55 TR (TPR value of 86.96% but OVA of 98.81%). Notwithstanding, if the majority class is also of equal or greater interest, the reverse choice might be preferable in which a model based on the 0.55 TR (TNR 99.68%, OVA 98.81%) would be adopted over that based on TR 0.81 (TNR 90.38%, OVA 91.04). Also see Tables 3.4a and 3.4c. Our study has clearly shown that the adapted PLS regression algorithm is adequate for handling chicken egg classification task. There is however a need to improve its present implementation mode, in relation to handling imbalanced data, towards achieving a better trade off between TPR and TNR accuracies, and at the same time favouring the use of lesser number of PLS components.

3.4 Conclusion

This chapter has presented the details of a study carried out to investigate the appropriateness of a PLS regression-based technique to classify chicken egg fertility data. Up to ten different set of features were extracted based on threshold selections. PLS regression (with an internal full cross validation procedure) analysis was implemented and tested for discrimination accuracy using the confusion matrix evaluation criterion. While 25 PCs were found suitable for accurate classification based on the true positive rates computation, only 5 PCs proved appropriate using the true negative and misclassification error rate computations. The analysis results showed that the adapted PLS regression algorithm can discriminate accurately both brown and white fertile

eggs from non-fertile eggs prior to incubation and on different days of incubation using the moving thresholding selection technique. It was further observed that recognising appropriately non-fertile eggs (TPR of 100%) with an acceptable matching up recognition accuracy of fertile eggs would need up to 25 PCs. However, models built with 5 PCs shifted recognition accuracies to be mostly in favour of the majority fertile egg class at the expense of the rare class non-fertile eggs. This scenario has been widely reported in literatures to be related to imbalanced data problem. We therefore need a classifier mode that will perform at much closer margin of TPR and TNR accuracies and at the same time using much lesser number of PCs. This study has clearly shown that the adapted PLS regression algorithm is adequate for handling chicken egg classification task, there is however a need to improve its present implementation mode, in relation to handling imbalanced data, towards achieving a better trade off between TPR and TNR accuracies, and at the same time optimizing the use of adequate number of PCs. Addressing the limitation in the present research outcome would be the major focus of our subsequent study.

CONNECTING TEXT TO CHAPTER 4

In chapter three, a supervised learning PLS regression approach was used to address chicken egg fertility classification task in its naturally imbalanced data structure. It was established that despite the adaptive PLS regression approach worked for the chicken egg classification task, the results were only promising with the use of many PLS components (PCs), as the use of fewer PCs shifted classification accuracies in favour of the prevalent class due to the imbalance data structure phenomenon. It is therefore imperative to improve on the present implementation mode of PLS for classification algorithm, in a view to accommodating the chicken egg fertility data structure and at the same time allowing the use of few and adequate number of PCs.

It was based on the results of chapter 3 that the second objective of this study (addressed in chapter 4) was drawn to examine the appropriateness of a PLSDA feature selection technique, towards improving model performance for chicken egg fertility discrimination. Chapter 4 has been submitted to the hyperspectral imaging laboratory of the Department of Bioresource Engineering, McGill University, for proprietary reason and awaiting approval for eventual submission to the Journal of IEE Transactions on Pattern Analysis and Machine Intelligence as:

Adegbenjo, A. O., and Ngadi, M. (2019). A non-parametric feature selection approach for chicken egg fertility classification using hyperspectral imaging. IEE Transactions on *Pattern Analysis and Machine Intelligence*

CHAPTER 4

A NON-PARAMETRIC FEATURE SELECTION TECHNIQUE FOR CHICKEN EGG FERTILITY CLASSIFICATION USING HYPERSPECTRAL IMAGING

Abstract

Our initial study has demonstrated the potential of a widely used PLS chemometric technique, for chicken egg fertility classification task. Our results are superb considering the whole data spectrum and use of relatively more PLS components (PCs). However, use of fewer PCs shifted our classification accuracies to be in favor of the majority class at the expense of the minority class, due to imbalance data structure. There is therefore a need for appropriate improvement in the adopted PLS for classification methodology. This sequel paper thereby presented the results of a partial least squares - discriminant analysis (PLSDA) based feature selection algorithm, with a non-parametric Receiver Operating Characteristic (ROC) curve analysis technique for identifying the most informative variables and thereby selecting appropriate discriminating features from chicken egg fertility data. Hyperspectral images total of 336 brown and 336 white hatchery eggs were captured in the NIR wavelength region of 900-1700 nm. Even though incubation period spans through day 0 until day 10, when all eggs were eventually broken out to confirm fertility, data recorded for subsequent downstream analysis was only for early incubation days 0-3. Spectral information was extracted from a segmented region of interest (ROI) of each hyperspectral image and a mean spectra matrix was afterwards obtained by averaging the spectral information. From a Monte-Carlo cross validation (MCCV) via a balanced subsampling, an optimal cut-off threshold was selected at center point 0.5 for identifying important discriminating features. Classification task was thereby simultaneously implemented using the same PLSDA algorithm. By considering only a maximum number of 5 PCs, classifier performance was measured from confusion matrix

and area under ROC curve (AUC) computations. Further verifications were also accomplished using nested cross validation, permutation testing, and "hold-out" set verification. With true positive rate (TPR), true negative rate (TNR), and AUC values in the range of 90-100% obtained on different days of incubation, the results indicated that the PLSDA based feature selection algorithm, in conjunction with a non-parametric ROC curve analysis technique can accurately identify discriminating features from chicken egg fertility data during early incubation, and thereby presented as suitable towards building an on-line fertility prediction system for chicken eggs using hyperspectral imaging.

4.1 Introduction

Over 50 billion of chickens are being raised annually by poultry farmers all over the world be it as layers, towards egg production or as broilers, towards meat production, and production growth is anticipated to continue. The global world population has been projected to hit 9.6 billion by 2050, creating an increasing demand for animal-based food (Mottet & Tempio, 2017). The Canadian chicken industry is a huge one with about 2,836 regulated producers spread across the provinces producing. There are about 241 broiler hatching egg producers and about 1,059 egg producers. Canada produced up to 1.2 billion kilograms of chicken in 2017, of which 61% of production originated from Quebec and Ontario (Agriculture and Agri-Food Canada, 2017). For the year 2017, Canada exported about 39.8 million hatching eggs, worth 68.8 million dollars as against 22.7 million hatching eggs in 2013 (worth $36 million). The importation figure for the same year stood at over 141 million hatching eggs for broilers (worth over $49 million), with entire importation coming from the US.

According to the (Agriculture and Agri-Food Canada report 2017), total hatching egg set (for both egg production chicks and broilers) was over 1.0 billion. With fertility rate observed in

the year 2017 to be around 82%, there were about 180 million unhatched eggs incubated in Canada for year 2017 alone. This meant a whooping sum of over 300 million Canadian dollars was wasted by the hatchery industries towards incubating unhatched eggs for the year 2017. Whereas, these non-hatching (non-fertile) eggs can find useful applications as commercial table eggs or low-grade food stock if they can be detected early and isolated accordingly preferably prior to incubation. Thus, identification and isolation of infertile eggs have notable economic benefits for commercial hatchery industries.

Furthermore, chicken egg fertility/early embryonic development detection belongs to the category of imbalanced data distribution during analysis. This is because the occurrence of fertile eggs is much more frequent than that of the non-fertile eggs. Indeed, it is known in the industrial settings and commercial hatcheries that up to 10% non-fertile eggs exist in any whole egg set batch. This occurrence was known to cause major setbacks on the classification accuracies of most existing learning algorithms (Liao, 2008; Sun et al., 2009), and so the need for more adequate approach for this specific kind of data.

Feature selection is a crucial step in machine learning procedure; especially when considering multidimensional data set like the hyperspectral imaging data. Feature selection aimed at selecting a subset of “n” features (“n” being a user-defined parameter) from an original set of “m” features, so that the feature space is optimally decreased in accordance to some evaluation criteria (Jason, 2016; Liu et al., 2014; Xia & Wishart, 2016). The selection of subsets per time to be learned are usually determined using the embedded, filter and /or wrapper methods (Ladha & Deepa, 2011; Saeys et al., 2007). While the filter approach considers the appropriateness of the selected features, independent of the classifying algorithm, the wrapper method on the other hand requires a classifier to evaluate feature appropriateness, but also can be computationally burdensome.

Whereas the filter techniques are classifier independent, simple and fast; they are limited due to their dark knowledge of the interaction between feature subset search and classifier (Liu et al., 2014). This disadvantage with the filter techniques is catered for in the wrapper and embedded methods. Also, there exist multivariate filters purposely developed to overcome limitation of the conventional filter approach, and these include the information gain, correlation and learner-based feature selection techniques (Hall, 1999; Jason, 2016; Liu et al., 2014). Even though feature selection has been an integral part of machine learning/data mining from inception, its capability towards resolving the rare class problem is only a recent eye opening (Hukerikar et al., 2011; Longadge & Dongre, 2013).

The PLS approach is a widely used technique that has found useful applications in various field of endeavours (Yu, 2000). Although the approach was not originally developed for statistical discrimination tasks, researchers have successfully used PLS for classification purposes (Barker & Rayens, 2003; Briandet et al., 1996; Gottfries et al., 1995; Iizuka & Aishima, 1997; Ortiz et al., 1996). In such situation, it is usually combined with linear discriminant analysis (LDA) as PLSDA to achieve classification objective rather than regression, when dependent Y variable is solely categorical (Pérez-Enciso & Tenenhaus, 2003). PLS has been reported to have capacity of surpassing PCA in performance, when within-groups variability dominates the among-groups variability (Barker & Rayens, 2003). In the specific case of an imbalance data situation as with the egg fertility data, the within-groups variability is also of a major concern in that there might exist several sub-clusters not having the same number of samples in a single class (Japkowicz, 2001). This existence of within-class imbalance is known to have detrimental effect on classifier performance (Yoon & Kwek, 2007), and therefore the need to address imbalance problem prior classification.

A receiver operating characteristics (ROC) curve is a well-known tool for evaluating the predictive performance of a binary classifier. Its use emerged from the field of signal detection, on to diagnostic systems in medicine, and now widely adopted in various other fields of endeavour (Egan, 1975; Fawcett, 2006; Swets, 1988; Swets et al., 2000). The earlier use of ROC in machine learning can be traced to the work of (Spackman, 1989), who showed the versatility of ROC curves in comparing and assessing algorithms. Its wide acceptance in machine learning was due to better understanding of the limitation of using only the conventional overall accuracy metric for measuring classifier performance (Provost & Fawcett, 1997; Provost et al., 1998). ROC curves possess additional characteristics that make them suitable for skewed data distribution and different classification error costs. These traits are of uttermost importance in research areas of cost-sensitive and imbalanced data learning (Fawcett, 2006), and so well suited for our purpose.

In view of the above, the present study has therefore considered and tested based on hyperspectral imaging, the suitability of partial least square discriminant analysis (PLSDA) based feature selection algorithm in conjunction with a non-parametric ROC curve analysis approach for identifying informative variables and thereby selecting appropriate discriminating features for chicken egg fertility classification. Performance evaluation have been based on the confusion matrix and area under ROC curve (AUC) computations. The adopted ROC analysis approach used with a PLSDA learner-based feature selection technique were considered advantageous over the more general parametric approach (Xia et al., 2013). Also, the confusion matrix evaluation criterion was considered more preferable over the conventional overall accuracy performance criterion which has been shown to be substandard especially when considered alone for data having skewed distribution (Liao, 2008).

4.2 Materials and Methods

4.2.1 Samples

A total of 336 Brown shell eggs and 336 White shell eggs were received from a commercial fertile egg producer (Simetin Hatchery; www.couvoir.com) in 14 batches (48 eggs per batch) over a period of 3 months. There were 7 batches of eggs collected in each group of brown and white egg sets. All eggs were first imaged by the hyperspectral imaging system on Day 0 (just prior to incubation) and immediately after imaging, the eggs were incubated in an Ova-Easy 190 Advance Series II Cabinet Incubator (Brinsea Products Inc., Florida, USA) at 37.78°C (100°F) and 55% relative humidity. The eggs were automatically turned every hour. On days 1, 2, and 3 of incubation, eggs were removed for imaging in sequence and then immediately returned into the incubator, in a process of about 1 min. After 10 days of incubation, eggs were broken out to determine fertility. Table 4.1 shows the details of the overall egg samples available for analysis on each day of incubation for both brown and white eggs. Out of the total 336 eggs received for both brown and white eggs, the number of total available eggs eventually used for analysis varied with incubation time due to egg breakage during handling. While 2 eggs (1, day 0; 1, day1) were broken from the brown egg batch, a total of 3 eggs (1, day 0; 2, day 3) were broken from the white egg batch. This variation in total available eggs during analysis results into a slightly different degree of imbalance from one day of incubation to another. The imbalance ratio of non-fertile to fertile eggs in this study is estimated from Table 4.1 to be 1:13 and 1:15 for both brown and white eggs respectively. This degree of imbalance according to (Sun et al., 2009) is enough to impede discrimination performance.

Table 4.1 Overall egg sample specifications for white and brown eggs

Incubation period	White Eggs					Brown Eggs				
	Egg received	Broken	Total Eggs used	Fertile (F)	Non-fertile (NF)	Egg received	Broken	Total Eggs used	Fertile (F)	Non-fertile (NF)
Day 0	336	1	335	314 (93.73%)	21 (6.27%)	336	1	335	312 (93.13%)	23 (6.87%)
Day 1	335	-	335	314 (93.73%)	21 (6.27%)	335	1	334	311 (93.11%)	23 (6.89%)
Day 2	335	-	335	314 (93.73%)	21 (6.27%)	334	-	334	311 (93.11%)	23 (6.89%)
Day 3	335	2	333	312 (93.69%)	21 (6.31%)	334	-	334	311 (93.11%)	23 (6.89%)

4.2.2 Image acquisition and processing

The hyperspectral imaging system used in this study consisted of an InGaAs camera, a line-scan spectrograph *(*HyperspecTM, Headwall Photonics Inc. USA) with a NIR spectral wavelength range from 900 to 1700 nm and a spectral resolution of 4.79 nm, a conveyor (Donner 2200 series, Donner Mfg. Corp., USA) driven by a stepping motor (MDIP22314, Intelligent motion system Inc., USA), a tungsten halogen lamp (50 W) providing back illumination to eggs, an enclosure supporting the system, a data acquisition and pre-processing software *(*Hyperspec, Headwall Photonics Inc. USA) and a PC. Spectral information was extracted from a segmented region of interest (ROI) of each hyperspectral image and a mean spectra matrix of 336 rows x 167 columns (samples x wavelength bands) was afterwards obtained by averaging the spectral information. Detailed image acquisition and processing procedure followed is as reported by (Liu & Ngadi, 2013).

4.2.3 Data resampling and pre-treatment

Apart from considering the entire batches of data together for analysis, the data were also resampled to monitor the effect of imbalance on analysis results. Data resampling is usually considered in terms of over-sampling minority class or under-sampling majority class. The approaches are usually implemented in a way to minimize any divergence between classes, thereby achieving a more uniform distribution of number of occurrences in each class (Liao, 2008; Nguyen et al., 2009). Due to overfitting disadvantage of oversampling (resulting from exact duplication of minority instances) and cogent information loss disadvantage of under-sampling the prevalent class, some more sophisticated approaches including but not limited to cluster based under-sampling and the “Synthetic Minority Oversampling Technique” (SMOTE) have been proposed (Chawla et al., 2002; Rahman & Davis, 2013; Yen & Lee, 2009). This present study followed a

modified cluster-based under-sampling approach, in which the data was first divided into two groups with one subgroup having all the prevalent (fertile egg) class examples and the other subgroup having the entire non-prevalent (non-fertile egg) class instances. The prevalent class examples were then further divided into "M" number of clusters (M >1), where each cluster is taken to be one subset of the prevalent class. All the subsets of the prevalent class were then separately combined with the rare (non-prevalent) class instances to make "M" different training data sets (the value of "M" is dependent on the data structure, "M" value is taken to be 7 in the present implementation). This cluster-based method has also been mentioned elsewhere as matched nested case-control study approach reported by (Xia et al., 2013). Further reading about matched nested case-control study and cluster-based resampling methodology can be obtained from literatures (Dunn et al., 2011; Dunn et al., 2012; Prachuabsupakij, 2015; Rahman & Davis, 2013; Rothman & Greenland, 1998; Yen & Lee, 2009). After the resampling procedure, both original raw and resampled data were further adjusted for batch to batch variability effect, using the empirical Bayes method as reported by (Johnson et al., 2007) and implemented in the MetaboAnalyst 3.0 platform.

Pre-treatment processes should be considered paramount during hyperspectral data handling because of high dimensionality phenomenon, multicollinearity, and existence of various unwanted variability that might be due to noise or low signal-to-noise ratio (ElMasry & Nakauchi, 2016). Common pre-treatment methods considered in this study include data integrity checking, missing value estimation, filtering, and normalisation. During integrity checking, data were checked for appropriate formatting, presence of dead pixels and missing/empty values. Features with greater than 50% missing values were automatically removed, and any other remaining missing values were estimated by small value replacement method computed as half of the

minimum positive value in the original raw data. Furthermore, initial dimensional reduction was accomplished via filtering (Hackstadt & Hess, 2009) using the interquartile range (IQR) estimation method. With this method, only 5% of the total 167 features considered were eliminated as redundant variables. Larger percentage of features can indeed be filtered out depending on the total number of features under consideration. Lastly, normalisation procedure was carried out using log transformation and auto scaling. Detailed explanation of the pre-treatment procedures carried out in this study is as described by (Xia & Wishart, 2016). All pre-treatments were carried out on MetaboAnalyst 3.0 platform.

4.2.4 Spectral transmission and feature selection

Spectral transmission characteristic namely Mean Spectral (MS) were further analysed in this study. MS stands for the mean value of all pixels in ROI for the current wavelength over the spectral range of 900-1700 nm. Understanding that simple models built with small numbers of variables rather than with a whole feature spectrum are usually less prone to overfitting, more robust, and likewise cost effective; this study have adopted a metabolomic feature selection approach for choosing important discriminating features prior a downstream multivariate analysis towards model development. In this approach, a PLSDA learner-based algorithm was evaluated on the data set considering various subsets of wavelength features, thereby selecting appropriate subset(s) that allows the classifier to achieve optimal performance. This procedure involved ranking the attributes used for model building in order of importance, and then repeating the whole modelling process using the top “k” features. Knowing that the ratio between two variable features might carry more discriminating information than two corresponding features considered individually (Xia & Wishart, 2016), top 20 ratio features were also computed and ranked with all other individual features during this feature selection stage of analysis. The learning and feature

ranking algorithms used were based on the PLSDA method in conjunction with a Monte-Carlo cross validation (MCCV) evaluation technique. An optimal cut off threshold centered at 0.5 was used in this multivariate analysis due to the balanced subsampling implementation mode of MCCV. The whole feature selection procedure was carried out on the MetaboAnalyst 3.0 software platform, using the biomarker identification module.

4.2.5 Partial least square discriminant analysis

For the different days of egg incubation, a PLSDA algorithm based on MCCV through balanced subsampling was implemented (Xia & Wishart, 2016; Xu & Liang, 2001). To apply MCCV in each analysis run, 70 % of total samples were used for important feature calibration and top 5, 10, 15, 25….100 (maximum) important features were selected and used to build suitable classification models, which were then validated on the left out 30% of the samples. This process was repeated a multiple of times and the average performance in terms of ROC curve was generated for each model. In the same vein, AUCs were computed and reported with their corresponding confidence intervals (CIs) (Hackstadt & Hess, 2009). A total of 840 discrimination models were built in this study (6 models for each batch, for each PC, and for each day of incubation) for both brown and white eggs respectively. However, only 4 models for each batch, for each PC, and for each day of incubation were considered in subsequent downstream analysis (these are models built with up to a maximum of 25 features), resulting in a total of 560 models considered in the final analysis for both brown and white egg sets respectively. Unlike the popular multiple linear regression (MLR) which is prone to the problem of over-fitting in the presence of too many factors, PLS analysis do adjust to this problem by extracting only the latent factors accounting for majority of the manifest factor variation (Tobias, 1995). Not only this, PLS is well known for analyzing data with strongly collinear (correlated), noisy, and numerous X-variables.

In the same vein, only 5 PLS components (PCs) in the maximum were considered throughout the analysis to minimize the effect of overfitting. This PCs functioned similarly like the principal components in principal component analysis (PCA). While PCA decomposes X to obtain components that explains most variability in X, PLS on the other hand seeks to identify components from X that best predict Y (Abdi, 2010).

4.2.6 Criteria for evaluating discrimination performance

Overall accuracy (considered alone) has been presented as inappropriate for measuring classifier performance in the situation consisting of a rare class data (Liao, 2008; Nguyen et al., 2009). The present study has therefore adopted the confusion matrix evaluation criterion for a binary-class egg fertility discrimination problem. ROC curves were further generated from the confusion matrix results and an area under ROC curve (AUC) values also computed with their corresponding confidence intervals. In a binary class classification situation, the particular class with very few training examples but with high identification importance is commonly referred to as the positive class and the other as the negative class (Sun et al., 2009). This definition has however been noted not to be directly applicable to most Agricultural and food processing operations. For example, non-fertile eggs in our study belonging to the rare class with very few training instances, are not of high recognition importance in comparison to the prevalent group fertile eggs. The present work and our subsequent studies have therefore adopted henceforth, the prevalent group (fertile egg class) as the positive class due to its high identification importance from the Industrial point of view. The hatchery industries are more interested in identifying fertile eggs as it seems costlier to them misclassifying fertile eggs than misclassifying non-fertile eggs. However, from algorithm development and implementation point of view, the reverse choice might be the case and this would be consistent with what is conventionally obtainable in various

other non-Agricultural/food applications, especially the biomedical and allied fields (Rahman & Davis, 2013; Sun et al., 2009; Teck et al., 2012).

The confusion matrix employed for the interpretation of the PLSDA analysis and hence determining classification accuracy is as shown in Table 4.2, where TPR, FPR, TNR, and FNR represent the rates in percentage of true positive, false positive, true negative, and false negative instances, respectively. If TP = True positive (number of fertile eggs correctly classified as fertile), TN = True negative (number of non-fertile eggs correctly classified as non-fertile), FP = False

Table 4.2 Confusion matrix

		Prediction class	
		Predicted as Positive	**Predicted as Negative**
True class	**Actual Positive**	TP	FN
	Actual Negative	FP	TN

positive (number of non-fertile eggs incorrectly classified as fertile), and FN = False negative (number of fertile eggs incorrectly classified as non-fertile), the following equations as reported by (Sun et al., 2009) can be obtained and the traditional overall accuracy (OVA) can also be computed:

$$TPR = TP/ (TP+FN) * 100 \quad 4.1$$

$$TNR = TN/(TN+FP) * 100 \quad 4.2$$

$$FPR = FP/ (FP+TN) * 100 \quad 4.3$$

$$FNR = FN/ (FN+TP) * 100 \quad 4.4$$

$$OVA = (TP+TN)/ (TP+FN+FP+TN) * 100 \quad 4.5$$

ROC curves were generated as scattered plot of TP (sensitivity) against FP (1-specificity). These curves, plotted multiples of times and averaged for smoothening, were then presented as a single metric called AUC. Utility of model features based on its AUC values (%) were then assessed as follows: 90-100 = excellent; 80-90 = good; 70-80 = fair; 60-70 = poor; and 50-60 = fail. Due to the limited data availability during laboratory studies compared to the eventual large population data expected for finalized model, there is always a need for CI computation. A 95% CI was adopted and calculated for all feature models presented. Even though the interpretation of CIs can be mostly subjective, great caution however must be taken to ascertain stability, parsimony and robustness of final feature model. Ideally, standard CI variability tolerance allowed for good model should not exceed 20 % (Xia et al., 2013). The wider the margin of the CI range, the greater the uncertainty being introduced in relation to the utility of the selected features.

4.2.7 Choice of optimal number of PLS components (PCs)

PLS modelling process is greatly influenced by only few underlying variables (latent variables); whereas, the appropriate number of these latent variables is usually unknown. One major aim of PLS analysis therefore was to estimate this number (Wold et al., 2001) and in doing so, it becomes very critical to identify an optimum value for the number of PCs which is directly related to the selection of the important variables/features required for accurate discrimination process.

The present study has followed a nested cross validation (CV) procedure (implemented within a MCCV loop) in testing the predictive performance of built models. From the CIs generated for the AUC results, standard errors of prediction (SE) were estimated. The SEs and CI ranges were then plotted against the number of PCs. Knowing that SE should ideally decrease as

more components are being added, optimal PCs were determined at the point where SE starts to increase upon adding an extra component. Adding more components above this optimal point is tantamount to overfitting. The CI range at this optimal point also should not ideally exceed 0.2 (Xia et al., 2013). A fuller detail of nested (double) CV and its other variations has been described elsewhere (Clark & Cramer, 1993; Filzmoser et al., 2009; Höskuldsson, 1988, 1996; Liebmann et al., 2010; Smit et al., 2007; Szymańska et al., 2012; Wakeling & Morris, 1993; Westerhuis et al., 2008; Wold et al., 1993).

4.2.8 Model verification

The final aim of any predictive modelling exercise is for the model to perform effectively when presented with unknown data. Hence, the number of PCs should neither be too small (underfitting) nor too big (overfitting). The only final litmus test to rule out these two loop holes is to have an independent evaluation of built models. This study therefore further tested three verification approaches namely repeated/nested CV prediction, permutation testing and unknown/hold-out sample prediction. In the hold-out set verification, 30% of the total data were left out unlabelled and used as testing set for built model. This testing set was randomly but carefully chosen to have appropriate representation of the training set. True performance of the built model was then tested from the hold-out set ROC curve analysis. It must be mentioned here that in an imbalanced data scenario, small samples or heterogenous sample populations, CV approach remains valid and its variant via MCCV subsampling and nested CV as earlier discussed and used in this study has been reported effective in generalising model's predictive ability to an unknown data set without necessarily creating a hold-out data set (Efron & Tibshirani, 1997; Eriksson et al., 2013; Picard & Cook, 1984; Xia et al., 2013). The second verification method called permutation testing is as described by (Good, 2011a, 2011b). This method is specifically

suited for ruling out the possibility of the optimal model built being because of random guess or chance. In this technique, the variable subset and model structure were fixed and multiple or randomly permuted models (e.g. N = 1000) were evaluated.

4.3 Results and Discussion

Tables 4.3 and 4.4 shows the evolution details of the egg samples in batches on each day of incubation for both original and resampled egg data respectively (where 'F' represents fertile eggs and 'NF' represents non-fertile eggs). Figure 4.1 shows typical preprocessing spectral transmission profiles for the egg data considered. The preprocessing profile patterns were similar irrespective of the incubation days (see Appendix B). Wherever indicated, label 0 stands for non-fertile eggs and label 1 stands for fertile eggs. It was observed that batch correction has no significant effect on the original raw egg data (Figure 4.1a and b). This observation is however more evident in the white egg set (Figure 4.1b) than in the brown egg set (Figure 4.1a). Whereas resampling operation showed some effects on the distribution pattern of the eggs (Figure 4.1c and d), batch correcting resampled egg data seems not to bring any additional improvement in the distribution pattern (Figure 4.1c and appendices B3 and B4).

Table 4.3 Original egg sample specifications in batches

	Brown Eggs (F, NF)				White eggs (F, NF)			
Days	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
Batch1	48(44,4)	48(44,4)	48(44,4)	48(44,4)	48(44,4)	48(44,4)	48(44,4)	48(44,4)
Batch2	48(45,3)	48(45,3)	48(45,3)	48(45,3)	47(43,4)	47(43,4)	47(43,4)	47(43,4)
Batch3	48(44,4)	48(44,4)	48(44,4)	48(44,4)	48(47,1)	48(47,1)	48(47,1)	46(45,1)
Batch4	48(46,2)	48(46,2)	48(46,2)	48(46,2)	48(44,4)	48(44,4)	48(44,4)	48(44,4)
Batch5	47(44,3)	46(43,3)	46(43,3)	46(43,3)	48(48,0)	48(48,0)	48(48,0)	48(48,0)
Batch6	48(43,5)	48(43,5)	48(43,5)	48(43,5)	48(42,6)	48(42,6)	48(42,6)	48(42,6)
Batch7	48(46,2)	48(46,2)	48(46,2)	48(46,2)	48(46,2)	48(46,2)	48(46,2)	48(46,2)
Total	335 (312,23)	334 (311,23)	334 (311,23)	334 (311,23)	335 (314,21)	335 (314,21)	335 (314,21)	333 (312,21)

Table 4.4 Resampled egg specifications in batches

	Brown Eggs (F, NF)				White eggs (F, NF)			
Days	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
Batch1	67(44,23)	67(44,23)	67(44,23)	67(44,23)	65(44,21)	65(44,21)	65(44,21)	65(44,21)
Batch2	68(45,23)	68(45,23)	68(45,23)	68(45,23)	64(43,21)	64(43,21)	64(43,21)	64(43,21)
Batch3	67(44,23)	67(44,23)	67(44,23)	67(44,23)	68(47,21)	68(47,21)	68(47,21)	**66(45**,21)
Batch4	69(46,23)	69(46,23)	69(46,23)	69(46,23)	65(44,21)	65(44,21)	65(44,21)	65(44,21)
Batch5	67(44,23)	**66**(**43**,23)	66(43,23)	66(43,23)	69(48,21)	69(48,21)	69(48,21)	69(48,21)
Batch6	66(43,23)	66(43,23)	66(43,23)	66(43,23)	63(42,21)	63(42,21)	63(42,21)	63(42,21)
Batch7	69(46,23)	69(46,23)	69(46,23)	69(46,23)	67(46,21)	67(46,21)	67(46,21)	67(46,21)
Total (rbc)	473 (312,161)	472 (311,161)	472 (311,161)	472 (311,161)	461 (314,147)	461 (314,147)	461 (314,147)	459 (312,147)

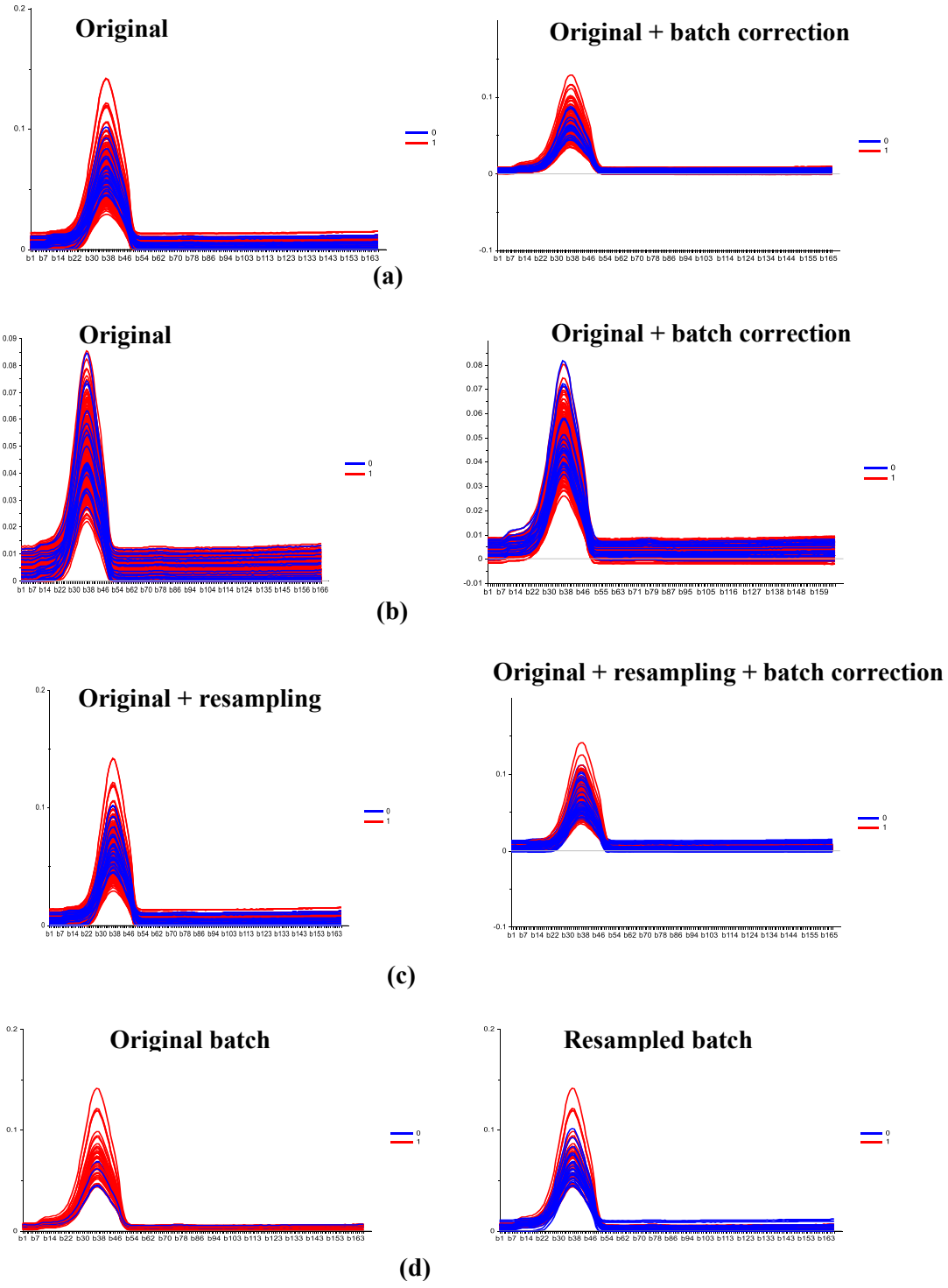


Figure 4.1: Typical preprocessing spectral transmission profiles (a), brown egg batch correction (b), white egg batch correction (c), resampling and batch correction (d), batch data resampling

It was seen from Figure 4.1c (appendix B3) and more clearly with individual batch data in Figure 4.1d (appendix B4), that resampling procedure exposes the egg data structure much better for subsequent analysis and learning. This phenomenon of data structure exposure to learning algorithms has always been a major consideration in spectra data preprocessing (Jason, 2016). The resulting resampled egg batch data are with noticeable overlapping bands due to multicollinearity spectral effect. This multicollinearity effect was later adjusted for via normalisation pre-treatment, thereby transforming the initial non-Gaussian batch data structure to a Gaussian like distribution data structure as depicted in Figure 4.2. This normalisation procedure remains very critical in hyperspectral data analysis since most classification algorithms including the PLSDA algorithm do assume a normal distribution data structure during implementation.

It was noted in a preliminary PCA and PLS analysis that the number of usable components in chicken egg hyperspectral data does not extend beyond the first 1 to 3 PCs. This understanding of potential informative features existing in chicken egg hyperspectral data, not exceeding those in the first 1 to 3 PCs was of uttermost consideration in the later choice of optimum number of PCs for final model building. From Figure 4.3 PCA and PLS loading plots, only the first 2PCs have explained 100% variability in the data and the onset of noise began from the 3rd PC. The PLS X-loadings further showed more evident noise progression in Factor 4 (PC4). It was from this observation that the optimal models chosen from this study were those built from components not exceeding the first 3PCs.

Figure 4.4 showed typical charts for determining the best model for each of the principal components considered and eventually determining the optimum PC. Models built at optimal PCs for each of the batches were thereby identified.

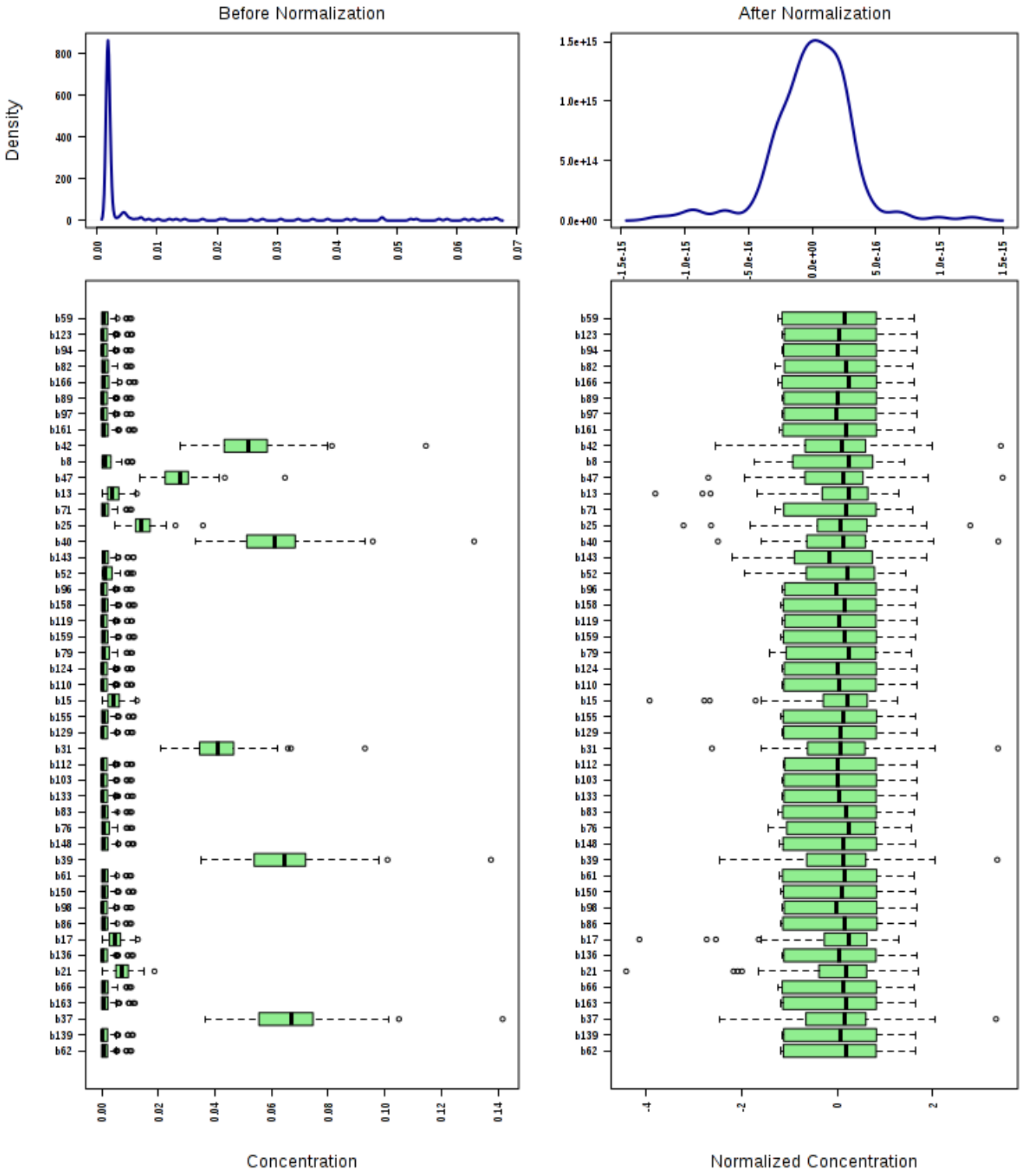


Figure 4.2: Typical egg batch spectral normalisation procedure

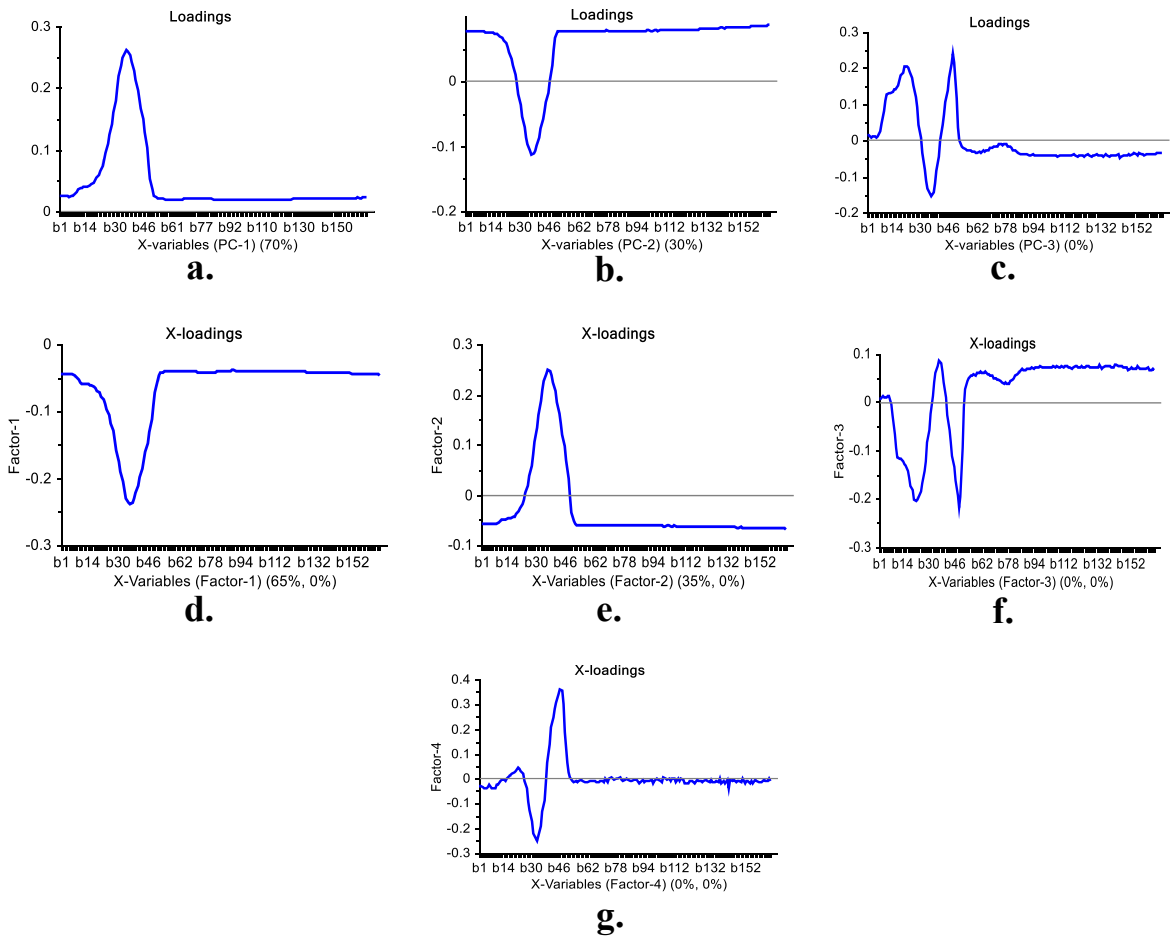
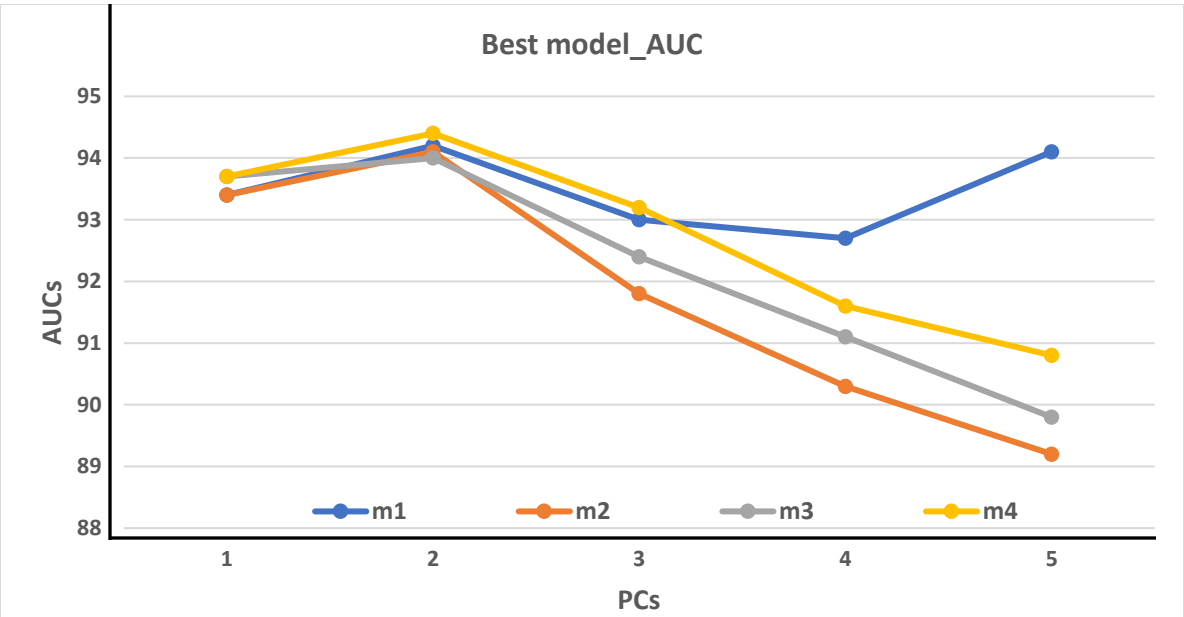
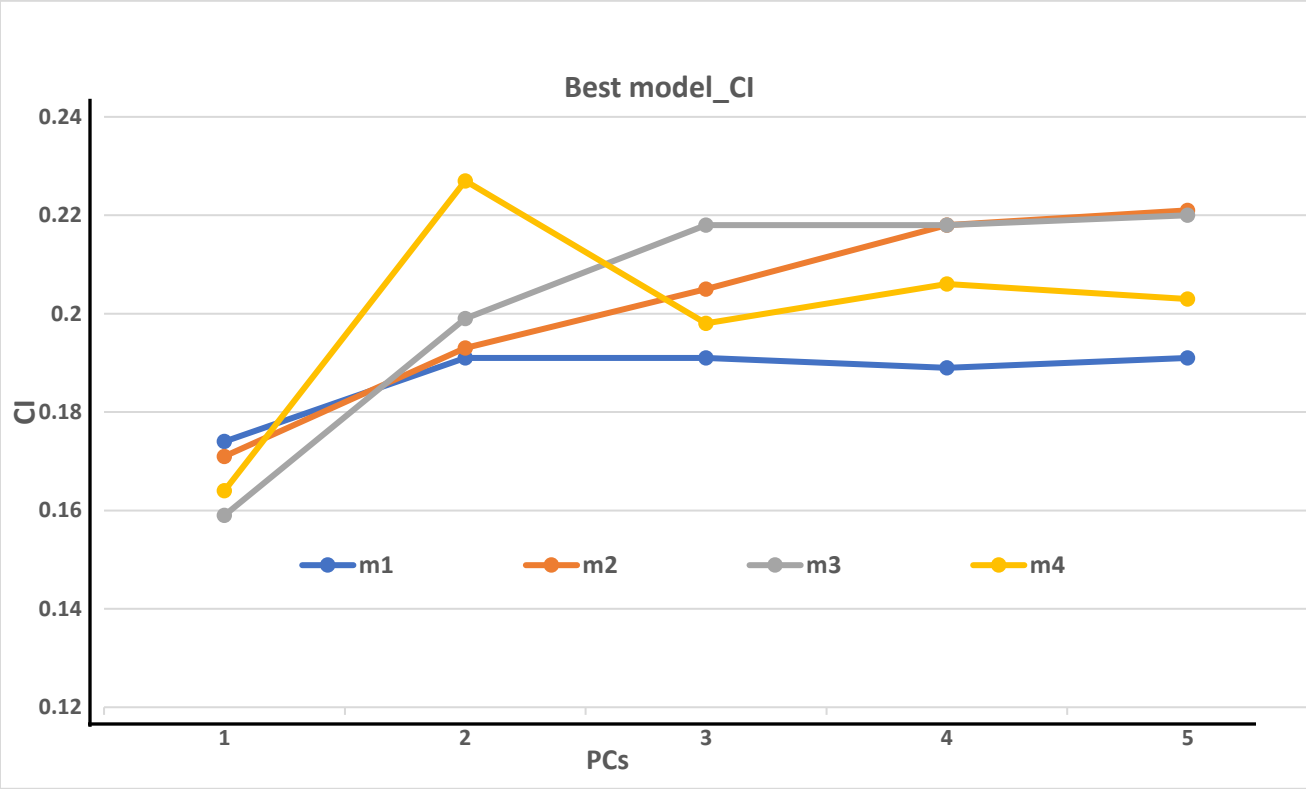


Figure 4.3: PCA and PLS loadings analysis plots (a), PCA 1st PC (b), PCA 2nd PC (c). PCA 3rd PC (d), PLS 1st factor (e), PLS 2nd factor (f), PLS 3rd factor (g), PLS 4th factor

From Figure 4.4a, models 3 and 4 at the first PC have the highest AUC value of 93.70%. However, considering the CI tolerances for the same set of models from Figure 4.4b, makes model 3 the optimal choice model with 0.159 CI tolerance against its model 4 counterpart having CI tolerance of 0.164. At PC2, model 4 has the highest AUC value of 94.40% (Figure 4.4a) but also the highest CI tolerance of 0.227. Knowing that any model with CI tolerance greater than 0.2 is unstable and possesses uncertainty utility of its selected features, model 1 with AUC value of 94.2% and CI tolerance of 0.191 was therefore chosen as optimal at PC2 (Figures 4.4a & b).



a.



b.

Figure 4.4: Selecting best model using values from (a), AUC (b) Confidence interval (CI)

Plotting CI tolerances and SE against number of PCs (Figure 4.5) for the best models earlier determined above (Figure 4.4), identified the first PC as optimum for this specific egg batch (brown egg batch 5). Hence, model 3 built with 15 features in the first PC becomes the final optimal chosen model for this batch (see the AUC model platform in Figure 4.6 for more clarity).

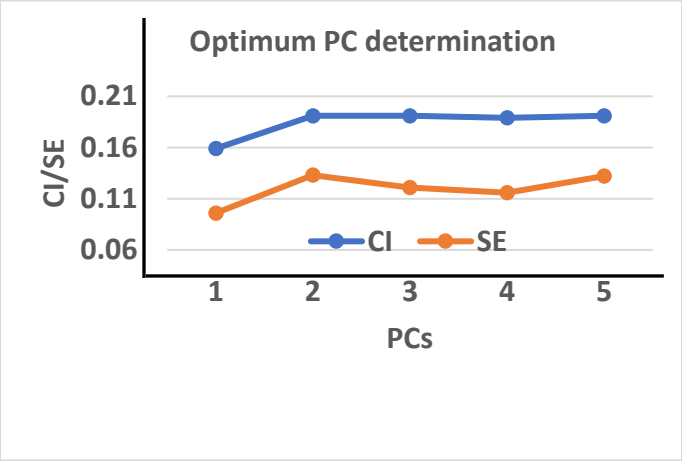


Figure 4.5: Selecting optimum number of PCs for a typical batch brown egg data

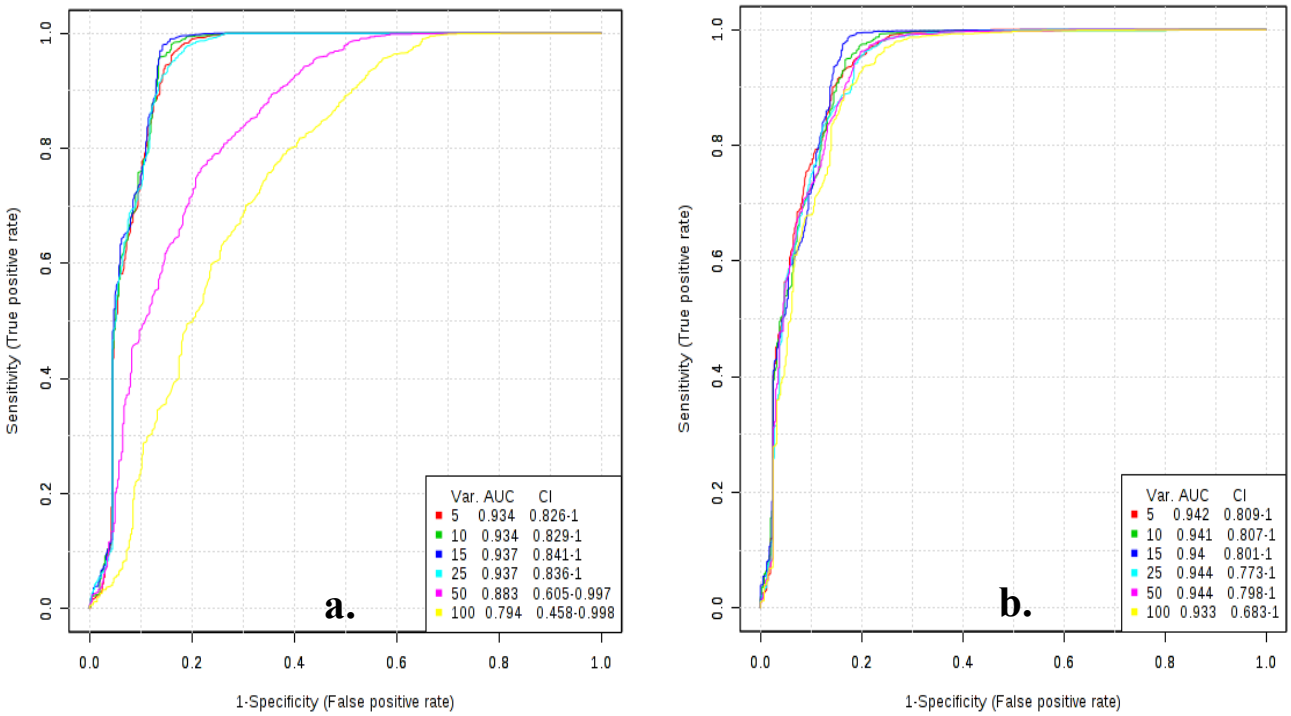


Figure 4.6: AUC model platforms for a brown egg batch 5 (a), PC1 models (b), PC2

Detailed choices of optimal models at different PCs for different days of incubation and for both brown and white eggs are as shown in Tables 4.5 and 4.6. It was observed from Tables 4.5 and 4.6 that batch by batch analysis of resampled data produced better results when compared to the analysis of overall raw egg data considered together. In the same vein, models built with feature ratio consideration outperform those built without feature ratio consideration. For all the days of incubation considered, best models without feature ratio consideration were built with up to 5PCs, whereas optimal PCs for best models built with ratio feature consideration do not extend beyond the first 3PCs. Furthermore, models built without ratio feature consideration are of a wider CI margins in comparison to models built with ratio feature consideration. This observation seems to be more evident in white eggs than in brown eggs (see Figure 4.7). For all days of incubation considered and for both brown and white eggs, optima models presented were mostly rated excellent possessing AUC values in the range of 90-100%. Hence, the PLSDA algorithm in conjunction with a non-parametric ROC curve analysis approach was confirmed appropriate for identifying important features towards early prediction of chicken egg fertility and thereby presented as a suitable algorithm towards building an online chicken egg fertility classification system, during early incubation.

Table 4.5 AUC values of best models with corresponding optimal PCs for brown eggs (a), day 0 (b), day1 (c), day 2 (d), day 3

		Non-ratio features						Ratio features					
a.	***PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	48.90 (1)	49.20 (2)	49.10 (1)	48.00 (2)	48.20 (1)	2	53.40 (1)	55.50 (1)	54.30 (1)	53.00 (1)	54.50 (1)	2
	Raw bc	46.70 (1)	48.70 (1)	50.00 (2)	51.20 (4)	50.00 (4)	4	62.40 (1)	65.50 (3)	62.50 (3)	60.90 (3)	63.10 (1)	2
	Raw rbc	40.90 (1)	56.10 (4)	61.00 (3)	63.10 (4)	63.50 (4)	5	61.00 (3)	65.60 (3)	65.30 (3)	65.50 (4)	66.00 (4)	5
	Batch1	69.00 (2)	71.8 (2)	73.00 (3)	76.00 (4)	78.00 (3)	5	86.10 (1)	86.10 (1)	86.00 (1)	86.10 (1)	84.50 (2)	1
	Batch2	60.50 (4)	64.80 (2)	67.20 (3)	78.20 (2)	80.10 (2)	5	**92.20 (3)**	92.80 (4)	90.80 (4)	87.30 (3)	91.80 (1)	1
	Batch3	82.30 (1,2,3)	78.30 (4)	79.90 (4)	77.60 (4)	79.40 (2)	1	88.90 (4)	89.00 (4)	89.50 (2)	87.90 (1)	91.30 (1)	5
	Batch4	93.60 (1)	**95.40 (4)**	95.50 (4)	95.00 (4)	94.40 (1)	2	93.70 (2)	94.70 (4)	**95.40 (3)**	95.60 (4)	96.50 (4)	3
	Batch5	40.80 (1)	47.70 (1)	57.90 (4)	61.70 (2)	69.70 (2)	5	**93.70 (3)**	94.20 (1)	93.00 (1)	92.70 (1)	94.10 (1)	1
	Batch6	58.00 (1)	65.20 (2)	87.30 (4)	90.60 (4)	90.70 (4)	5	**96.20 (4)**	94.40 (4)	94.80 (4)	91.60 (4)	92.20 (1)	1
	Batch7	64.90 (3)	73.80 (3)	79.70 (2)	82.10 (2)	84.40 (3)	5	96.20 (1)	**96.70 (3)**	95.50 (4)	95.00 (1)	95.90 (1)	2
b.	**PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	47.90 (4)	45.30 (3)	49.10 (3)	48.00 (2)	46.50 (2)	3	56.60 (1)	65.50 (2)	63.90 (1)	62.70 (1)	63.20 (1)	2
	Raw bc	47.00 (1)	46.80 (1)	49.50 (1)	51.80 (1)	48.40 (2)	4	52.20 (3)	55.40 (3)	50.00 (1)	50.50 (1)	50.00 (1)	2
	Raw rbc	50.60 (1)	51.00 (2)	53.00 (1)	57.00 (1)	58.50 (2)	5	65.80 (3)	68.00 (4)	67.90 (4)	66.90 (2)	66.70 (3)	2
	Batch1	50.80 (2)	58.20 (1)	69.80 (1)	78.00 (2)	86.30 (4)	5	**98.40 (3)**	97.10 (4)	94.90 (3)	94.50 (3)	95.30 (1)	1
	Batch2	45.30 (1)	49.50 (3)	57.60 (1)	71.20 (3)	77.20 (2)	5	96.40 (4)	96.20 (4)	**96.30 (4)**	96.20 (4)	95.80 (4)	3
	Batch3	50.20 (1)	52.20 (3)	58.10 (4)	64.80 (3)	73.50 (4)	5	94.30 (3)	**97.00 (4)**	95.30 (4)	94.60 (1)	94.90 (1)	2
	Batch4	45.00 (2)	55.80 (4)	74.10 (4)	79.50 (2)	83.50 (4)	5	93.50 (2,3,4)	95.00 (4)	**95.70 (3)**	95.30 (3)	94.70 (3)	3
	Batch5	54.70 (1)	53.90 (4)	61.20 (3)	75.20 (3)	85.60 (4)	5	**91.70 (4)**	91.80 (4)	94.90 (4)	94.80 (4)	94.20 (4)	1
	Batch6	73.70 (4)	70.80 (3)	74.40 (2)	78.20 (3)	80.60 (3)	5	88.10 (1)	85.80 (4)	84.50 (1)	85.00 (1)	85.20 (1)	1
	Batch7	84.00 (1)	82.50 (4)	83.90 (1)	91.60 (3)	93.50 (3)	5	**99.10 (4)**	98.20 (4)	98.30 (4)	97.50 (4)	97.80 (1)	1
c.	**PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	48.50 (1)	47.60 (1)	47.00 (1)	47.30 (1)	46.60 (1)	1	62.90 (1)	60.20 (3)	59.30 (1)	57.80 (1)	58.70 (1)	1
	Raw bc	44.20 (1)	46.60 (1)	45.00 (1)	44.10 (2)	45.50 (3)	5	60.00 (2)	59.90 (1)	59.10 (1)	57.90 (1)	60.00 (1)	1
	Raw rbc	50.00 (1)	55.80 (3)	56.60 (2)	58.20 (2)	62.40 (3)	5	68.80 (1)	71.00 (3)	72.60 (3)	75.00 (4)	75.4 (4)	5
	Batch1	65.40 (2)	78.20 (2)	79.60 (4)	85.30 (2)	86.10 (3)	5	91.30 (3)	93.10 (1)	**93.00 (1)**	92.60 (2)	92.40 (2)	3
	Batch2	50.90 (2)	50.20 (3)	64.60 (1)	73.80 (2)	81.70 (4)	5	94.20 (3)	**95.00 (4)**	93.90 (3)	94.30 (1)	94.20 (1)	2
	Batch3	50.10 (4)	55.70 (2)	66.10 (2)	78.00 (2)	83.60 (2)	5	94.80 (4)	**96.60 (3)**	96.00 (3)	95.40 (2)	94.90 (2)	2
	Batch4	81.80 (3)	83.40 (4)	83.90 (1)	85.00 (1)	87.80 (2)	5	94.70 (2)	**96.40 (4)**	96.60 (4)	95.90 (1)	96.50 (1)	2
	Batch5	50.10 (1)	50.00 (1)	66.60 (1)	78.60 (2)	82.60 (2)	5	**92.80 (3)**	91.60 (2)	91.00 (3)	92.00 (1)	92.00 (1)	1
	Batch6	53.60 (1)	54.60 (2)	68.10 (2)	77.60 (2)	83.20 (3)	5	93.20 (3)	**94.30 (4)**	92.40 (1,3)	92.20 (1)	92.10 (1)	2
	Batch7	74.00 (1)	77.50 (2)	78.40 (1)	93.10 (2)	95.20 (3)	5	**97.20 (4)**	96.90 (3)	96.50 (4)	95.80 (4)	95.90 (1)	1
d.	**PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	57.70 (1)	55.20 (2)	54.30 (1)	53.50 (1)	52.90 (1)	1	63.00 (2)	61.30 (3)	59.00 (4)	57.00 (1)	58.60 (1)	1
	Raw bc	57.60 (4)	50.20 (1)	49.30 (2)	48.10 (4)	52.60 (4)	1	63.40 (3)	59.10 (1)	58.40 (2)	53.40 (1)	58.70 (1)	1
	Raw rbc	61.30 (1,2)	62.20 (2)	63.60 (4)	68.70 (4)	70.80 (2)	5	64.30 (2)	64.00 (4)	64.20 (2)	65.50 (4)	65.60 (4)	5
	Batch1	72.60 (1)	74.90 (1)	80.00 (2)	80.40 (2)	80.90 (4)	5	**94.80 (4)**	93.10 (4)	93.00 (1)	92.90 (4)	93.80 (1)	1
	Batch2	63.70 (1)	65.20 (4)	61.90 (1)	62.00 (1)	69.90 (3)	5	95.20 (3)	**95.60 (4)**	95.40 (4)	94.20 (4)	94.90 (1)	2
	Batch3	50.30 (1)	57.60 (2)	63.40 (4)	72.50 (3)	79.50 (4)	5	**95.00 (3)**	92.80 (1)	91.80 (2)	92.30 (1)	93.00 (1)	1
	Batch4	76.10 (1)	76.00 (3)	74.90 (2)	75.30 (2)	79.20 (2)	5	96.20 (3)	**96.20 (4)**	96.00 (3)	96.10 (3)	96.30 (1)	2
	Batch5	77.50 (4)	78.80 (1)	82.00 (1)	83.50 (2)	84.10 (3)	5	93.40 (1)	91.60 (2)	**91.00 (3)**	92.00 (1)	92.00 (1)	3
	Batch6	51.40 (4)	55.30 (3)	62.00 (1)	67.80 (2)	69.80 (2)	5	90.50 (3)	89.60 (4)	87.20 (1)	86.10 (4)	87.70 (1)	1
	Batch7	93.00 (2)	93.70 (4)	94.30 (1)	98.30 (1)	97.30 (3)	4	**96.90 (4)**	95.80 (4)	96.10 (4)	96.40 (4)	95.80 (4)	1

* AUC values in this table are displayed for each of the PCs with model numbers placed in bracket.

Table 4.6 AUC values of best models with corresponding optimal PCs for white eggs (a), day 0 (b), day1 (c), day 2 (d), day 3

		Non-ratio features						Ratio features					
a.	***PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	42.50 (1)	44.60 (4)	46.00 (3)	48.30 (3)	47.20 (3)	4	59.60 (2)	61.10 (1)	61.00 (1)	59.60 (1)	59.90 (1)	2
	Raw bc	47.20 (1)	49.30 (3)	49.10 (2)	50.80 (3)	51.60 (2)	5	62.70 (2)	64.60 (3)	64.30 (4)	62.10 (4)	60.00 (4)	2
	Raw rbc	47.00 (4)	51.40 (3)	61.20 (4)	62.90 (4)	61.70 (3)	4	71.00 (1)	75.10 (2)	76.10 (3)	75.70 (3)	74.90 (2)	3
	Batch1	72.50 (3)	73.20 (3)	75.50 (3)	76.60 (4)	79.00 (4)	5	92.40 (2)	**95.50 (3)**	94.30 (4)	91.40 (4)	93.10 (1)	2
	Batch2	60.50 (4)	65.20 (3)	70.20 (3)	76.50 (3)	77.10 (3)	5	91.90 (1)	**94.20 (3)**	91.90 (3)	90.30 (1)	92.40 (1)	2
	Batch3	69.10 2)	72.70 (4)	76.60 (4)	88.10 (2)	90.30 (2)	5	97.80 (2)	**99.00 (4)**	98.30 (3)	98.30 (4)	98.30 (4)	2
	Batch4	80.10 (3)	78.40 (3)	75.50 (4)	77.20 (4)	76.90 (1)	1	87.00 (3)	88.00 (4)	86.40 (2)	84.80 (3)	87.90 (1)	5
	Batch5	48.60 (1)	53.40 (4)	71.80 (4)	78.50 (3)	83.60 (2)	5	100.0(1)	99.80 (1)	**99.90 (4)**	99.50 (1)	99.10 (2)	3
	Batch6	71.30 (3)	69.50 (1)	68.60 (3)	71.00 (1)	72.00 (3)	5	87.30 (4)	86.00 (4)	**88.40 (4)**	85.10 (4)	86.60 (1)	3
	Batch7	77.00 (2)	79.80 (1)	86.80 (1)	92.90 (2)	93.90 (2)	5	90.90 (2)	**92.60 (4)**	92.10 (3)	92.10 (1)	92.50 (1)	2
b.	**PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	58.10 (3)	54.10 (4)	58.50 (4)	57.50 (2)	54.70 (2)	3	68.90 (2)	66.10 (2)	64.80 (1)	62.60 (1)	61.80 (1)	1
	Raw bc	48.10 (3)	55.10 (3)	57.40 (1)	58.50 (3)	58.00 (2)	4	65.50 (3)	65.00 (2)	63.90 (2)	61.00 (3)	61.80 (1)	1
	Raw rbc	43.00 (3,4)	56.90 (4)	68.60 (2)	70.20 (1)	75.70 (3)	5	64.30 (1)	68.70 (2)	74.60 (4)	74.90 (4)	75.30 (4)	5
	Batch1	59.10 (4)	65.70 (2)	72.10 (1)	81.00 (2)	83.30 (2)	5	**95.10 (4)**	92.00 (3)	92.40 (3)	90.60 (1)	91.40 (1)	1
	Batch2	55.80 (2)	62.80 (1)	66.30 (3)	81.10 (3)	83.80 (4)	5	**93.90 (3)**	93.50 (1)	93.00 (2)	93.20 (1)	93.40 (1)	1
	Batch3	98.30 (2,3,4)	98.70 (4)	98.20 (4)	97.60 (4)	98.50 (1)	2	**98.60 (1)**	98.80 (4)	98.40 (4)	97.90 (4)	98.50 (1)	1
	Batch4	67.20 (1)	66.60 (2)	66.70 (2)	72.50 (2)	75.30 (3)	5	90.00 (3)	89.80 (1)	87.90 (3)	85.40 (3)	89.90 (1)	5
	Batch5	84.00 (2)	85.50 (1)	95.20 (1)	98.80 (2)	99.30 (3)	5	98.30 (4)	99.50 (4)	**99.80 (4)**	99.50 (4)	99.20 (4)	3
	Batch6	57.00 (4)	62.50 (3)	69.40 (4)	75.10 (3)	72.00 (2)	4	83.70 (4)	81.70 (3)	78.90 (4)	76.50 (4)	73.90 (3)	1
	Batch7	74.20 (1)	84.00 (2)	84.60 (1)	93.60 (3)	94.50 (3)	5	**96.80 (3)**	96.20 (4)	95.30 (2)	95.20 (4)	94.80 (2)	1
c.	**PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	48.80 (4)	48.40 (2)	47.00 (1)	51.90 (2)	54.80 (4)	5	69.50 (1)	71.20 (4)	69.10 (2)	66.90 (3)	67.40 (4)	2
	Raw bc	44.00 (3)	52.30 (2)	50.20 (2)	52.80 (2)	56.00 (4)	5	75.80 (3)	78.60 (3)	75.70 (2)	74.50 (2)	74.80 (1)	2
	Raw rbc	49.40 (2)	55.40 (4)	58.20 (4)	62.40 (2)	67.30 (3)	5	77.70 (3)	78.10 (3)	79.20 (4)	80.10 (4)	79.20 (4)	4
	Batch1	79.60 (1)	80.00 (1)	78.60 (2)	78.30 (1)	79.10 (1)	2	90.10 (1)	92.80 (1)	**92.80 (2)**	92.30 (1)	93.40 (1)	2
	Batch2	59.20 (4)	73.40 (3)	75.70 (1)	84.70 (1)	85.80 (2)	5	94.60 (4)	95.60 (4)	**92.60 (4)**	92.00 (1)	92.50 (1)	2
	Batch3	52.50 (5)	57.60 (2)	67.00 (1)	82.30 (3)	91.60 (4)	5	99.20 (4)	99.30 (4)	**99.30 (4)**	99.10 (4)	98.90 (4)	2
	Batch4	59.50 (3)	60.50 (1)	66.80 (2)	80.00 (3)	85.70 (3)	5	**93.40 (3)**	92.50 (4)	93.10 (4)	90.60 (4)	88.60 (4)	1
	Batch5	60.30 (1)	65.90 (2)	77.20 (1)	89.80 (2)	95.20 (3)	5	**100.00 (1,2,3,4)**	**100.00 (1,2,3,4)**	**100.00 (1,2,3,4)**	**100.00 (1,2,3,4)**	**100.00 (1,2,3,4)**	-
	Batch6	47.00 (2)	56.30 (4)	65.90 (2)	74.60 (2)	74.50 (3)	4	91.40 (3)	90.20 (3)	89.00 (3)	86.40 (3)	84.60 (3)	1
	Batch7	63.70 (2)	85.50 (3)	82.30 (2)	85.40 (3)	90.00 (4)	5	99.20 (2)	99.20 (4)	98.40 (1)	98.30 (1)	98.60 (1)	1,2
d.	**PCs (m)**	**1**	**2**	**3**	**4**	**5**	**OPC**	**1**	**2**	**3**	**4**	**5**	**OPC**
	Raw	59.70 (4)	57.0 (2)	55.70 (3)	58.20 (4)	55.60 (4)	1	67.30 (4)	67.10 (2)	62.80 (1)	62.10 (1)	62.30 (1)	2
	Raw bc	46.50 (1)	57.90 (4)	57.20 (2)	59.00 (2)	60.70 (4)	5	61.60 (2)	63.10 (4)	63.00 (4)	61.00 (4)	61.60 (4)	2
	Raw rbc	58.70 (4)	63.50 (3)	76.60 (4)	78.10 (4)	79.80 (4)	5	70.60 (4)	76.20 (4)	77.60 (4)	78.50 (4)	79.40 (4)	5
	Batch1	49.40 (4)	53.50 (4)	66.80 (4)	78.60 (2)	82.60 (3)	5	**96.10 (3)**	94.40 (1)	94.10 (1)	94.20 (1)	94.20 (1)	1
	Batch2	95.40 (1)	95.90 (1)	94.20 (2)	94.00 (2)	95.70 (2)	5	**96.80 (4)**	96.50 (4)	94.10 (2)	94.30 (1)	95.00 (1)	1
	Batch3	88.00 (3)	89.60 (1)	89.70 (2)	93.00 (2)	94.10 (3)	5	98.00 (4)	98.70 (1)	98.20 (1)	98.30 (4)	98.60 (1)	2
	Batch4	53.50 (1)	54.30 (1)	61.30 (1)	73.50 (4)	79.30 (3)	5	92.30 (3)	92.20 (4)	88.40 (4)	85.20 (3)	84.70 (4)	2
	Batch5	64.80 (1)	68.80 (1)	79.80 (1)	91.00 (2)	97.00 (4)	5	100.00 (1,2,3,4)	99.90 (1)	**99.90 (1)**	99.90 (1)	100.00 (1,2,3,4)	3
	Batch6	55.70 (4)	51.70 (4)	63.10 (1)	78.20 (2)	81.90 (4)	5	**92.90 (4)**	89.70 (4)	86.90 (2)	82.40 (4)	89.80 (1)	1
	Batch7	70.90 (1)	74.60 (1)	90.00 (1)	95.90 (3)	97.80 (3)	5	99.30 (3)	**99.30 (4)**	98.50 (4)	98.10 (4)	97.80 (4)	2

* AUC values in this table are displayed for each of the PCs with model numbers placed in brackets.

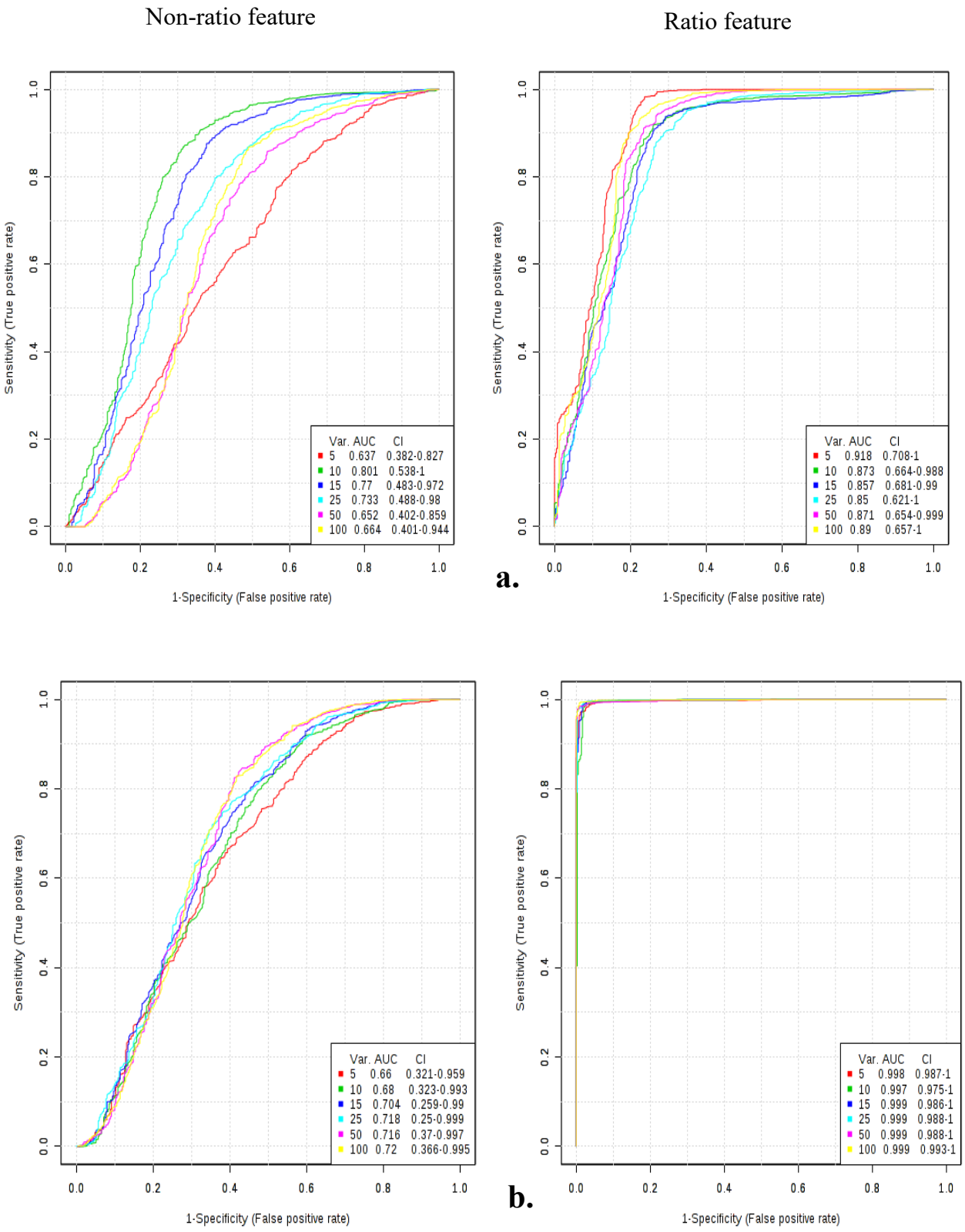
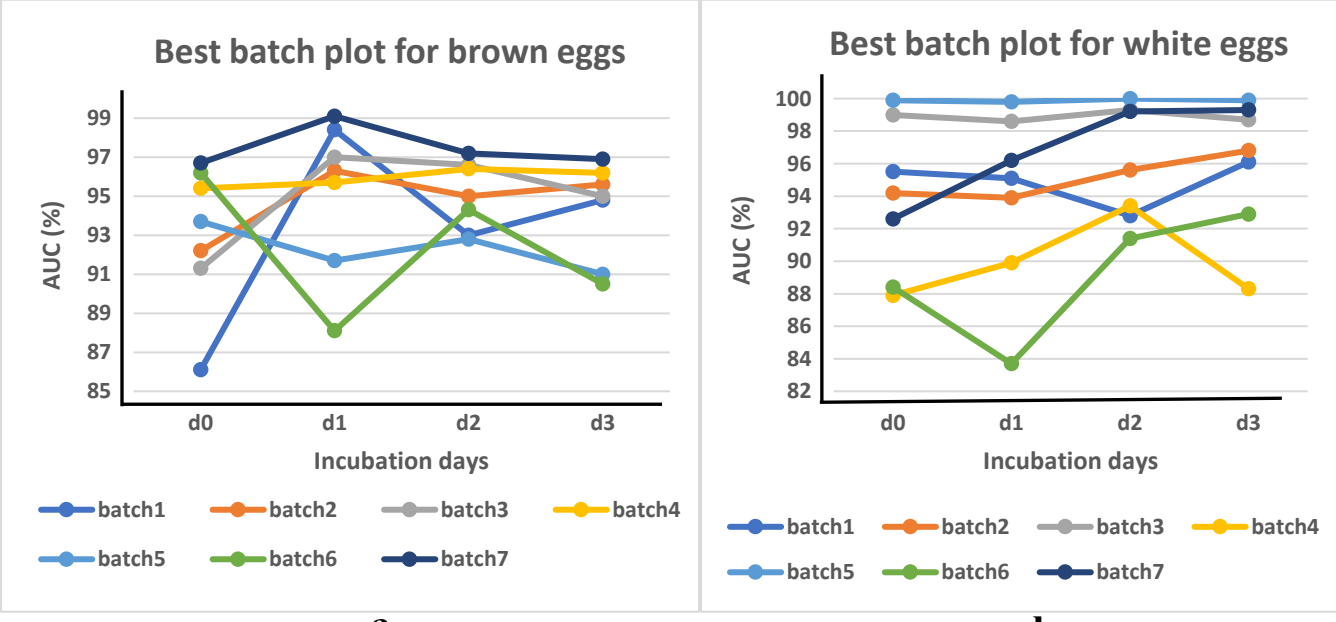


Figure 4.7: Non-ratio and ratio AUC feature model platforms (a), brown, batch 2 PC5 models (b), white, batch 5 PC3 models

Figure 4.8 shows for both brown and white eggs, the plots of best model AUCs against incubation days for all batches of eggs considered. From Figures 4.8a and 4.8b, Table 4.6 was extracted showing the least, median and best performing batch models on each day of incubation, and for both brown and white eggs. Some models have been rejected based on their confidence interval tolerance exceeding the maximum tolerance limit of 20% or just too close to zero. For example, on day 0 for white eggs; the least, median and best performing batch models ideally should be batches 4 (87.90%), 2 (94.20%), and 5 (99.90%) models respectively. However, knowing that batches 4, 6, and 7 models failed CI tolerance testing on this day, made the optimal models of choice to be batches 2 (94.20%), 1 (95.5%), and 5 (99.90%) for the least, median and best performing batch models respectively. This procedure was followed through all days of incubation to come out with Tables 4.7 and 4.8 optimal batch models for both brown and white eggs, which were later used for model verification in terms of repeated CV prediction, permutation testing, and hold-out set verification.



a. **b.**

Figure 4.8: AUC model plots for identifying optimal batch models (a), brown (b), white eggs

Table 4.7. Brown eggs **e**xtracted **o**ptimal batch models for each day of incubation
***(pc, m)**

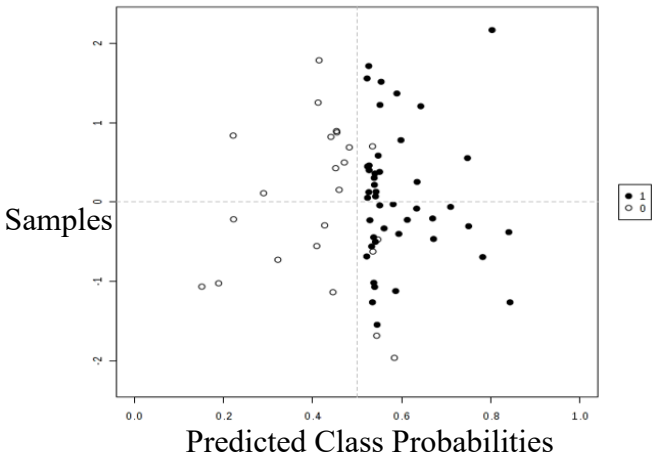
Incubation	Least batch	Median batch	Best batch	Excluded	Total Excluded
Day 0	Batch 2 (1,3)	**Batch 4 (3,3)**	**Batch 7 (2,3)**	1, 3	2
Day 1	Batch 5 (1,4)	Batch 4 (3,3)	Batch 7 (1,4)	6	1
Day 2	Batch 5 (1,3)	Batch 2 (2,4)	Batch 7 (1,4)	-	-
Day 3	Batch 5 (3,3)	Batch 1 (1,4)	Batch 7 (1,4)	6	1

Table 4.8. White eggs **e**xtracted **o**ptimal batch models for each day of incubation
***(pc, m)**

Incubation	Least batch	Median batch	Best batch	Excluded	Total Excluded
Day 0	**Batch 2 (2,3)**	**Batch 1 (2,3)**	**Batch 5 (3,4)**	4, 6, 7	3
Day 1	Batch 2 (1,3)	Batch 1 (1,4)	Batch 5 (3,4)	4, 6	2
Day 2	Batch 4 (1,3)	Batch 2 (2,4)	Batch 3 (2,4)	1, 5, 6	3
Day 3	Batch 6 (1,4)	Batch 1 (1,3)	Batch 7 (2,4)	4, 5	2

***(pc, m)** = (principal component, and model numbers)

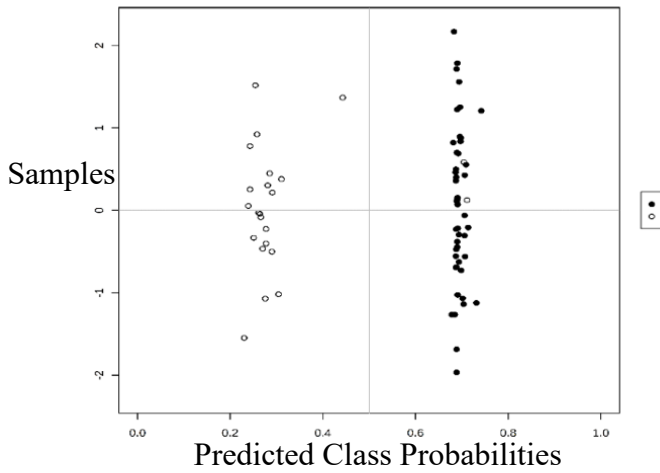
Figures 4.9 showed the MCCV classification and confusion matrix results for the identified optimal batch models on day 0 incubation for brown eggs. It was noted from Figure 4.8a, and Table 4.6a that despite the best model (batch7, day 0) performs better than the median model (batch4, day 0) in terms of AUC values (96.7% versus 95.40%), the median model (Figure 4.9b) on the other hand outperforms the best model (Figure 4.9c) in terms of classification accuracy (97.10% versus 94.20%) and confusion matrix results (see TPR, TNR, FPR, and FNR values from Figures 4.9 b and c). This observation might be due to the median model being built with 3PCs while the best model was built with only 2PCs (Table 4.6a). The third PC in the median model had seemingly contributed additional distinguishing feature to the model.



a.

True class	Prediction class: Predicted Positive	Prediction class: Predicted Negative
Actually Positive	100 (45/45)	0 (0)
Actually Negative	21.74 (5/23)	78.26 (18/23)

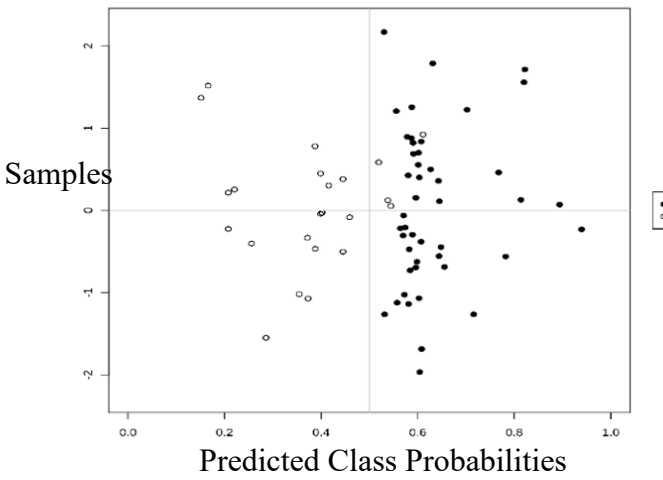
OVA = 92.65%



b.

True class	Prediction class: Predicted Positive	Prediction class: Predicted Negative
Actually Positive	100 (46/46)	0 (0)
Actually Negative	8.70 (2/23)	91.30 (21/23)

OVA = 97.10%



c.

True class	Prediction class: Predicted Positive	Prediction class: Predicted Negative
Actually Positive	100 (46/46)	0 (0)
Actually Negative	17.39 (4/23)	82.61 (19/23)

OVA = 94.20%

Figure 4.9: Brown eggs MCCV classification accuracy and confusion matrix results on day 0 incubation (a), least batch (b), median batch (c), best batch models

Some parameter details and results of confusion matrix, nested CV and permutation testing for both brown and white eggs were as shown in Tables 4.9 and 4.10 respectively.

Table 4.9. Training and testing prediction results for brown eggs

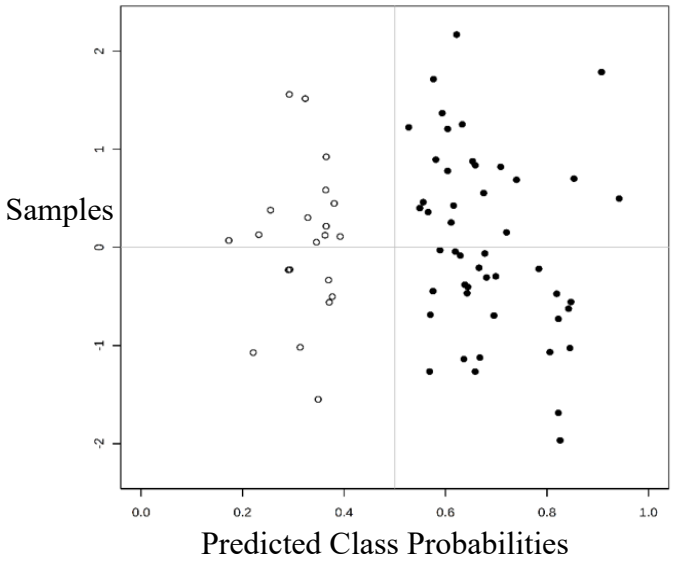
	Monte-Carlo CV						Nested CV and Permutation						
	AUC (%)	TPR (%)	TNR (%)	CLA (%)	MCVA (%)	CI %	AUC (%)	TPR (%)	TNR (%)	CLA (%)	NCVA (%)	CI %	P < 0.001
Day 0													
B2	92.20	100.00	78.26	92.65	93.00	15.00	94.90	100.00	86.96	95.59	95.10	12.70	PASS
B4	95.40	100.00	91.30	97.10	98.30	10.90	95.80	100.00	91.30	97.10	98.30	12.30	PASS
B7	96.70	100.00	82.61	94.20	96.30	10.00	98.30	100.00	82.61	94.20	96.30	07.30	PASS
Day 1													
B5	91.70	100.00	86.96	95.45	96.60	19.10	93.00	100.00	86.96	95.45	97.30	18.10	PASS
B4	95.70	100.00	91.30	97.10	95.50	10.10	96.40	100.00	91.30	97.10	96.60	11.70	PASS
B7	99.10	100.00	91.30	97.10	98.20	03.20	98.70	100.00	91.30	97.10	98.30	04.80	PASS
Day 2													
B5	92.80	100.00	86.96	95.45	96.70	16.40	94.80	100.00	86.96	95.45	97.30	13.20	PASS
B2	95.00	100.00	86.96	95.59	97.30	13.80	95.40	100.00	86.96	95.59	97.00	11.50	PASS
B7	97.20	100.00	91.30	97.10	98.20	08.10	96.90	100.00	91.30	97.10	98.30	10.90	PASS
Day 3													
B5	91.00	100.00	82.61	93.94	90.80	16.90	92.20	97.67	82.61	92.42	91.20	18.30	PASS
B1	94.80	97.73	73.91	89.55	92.30	13.30	95.90	100.00	82.61	94.03	96.00	11.90	PASS
B7	96.90	100.00	82.61	94.20	93.90	10.80	96.50	100.00	82.61	94.20	97.20	11.70	PASS

Table 4.10. Training and testing prediction results for white eggs

	Monte-Carlo CV						Nested CV and Permutation						
	AUC (%)	TPR (%)	TNR (%)	CLA (%)	MCVA (%)	CI %	AUC (%)	TPR (%)	TNR (%)	CLA (%)	NCVA (%)	CI %	P < 0.001
Day 0													
B2	94.20	100.00	71.43	90.62	92.10	15.70	93.30	100.00	71.43	90.62	93.70	15.80	PASS
B1	95.50	100.00	80.95	93.84	94.60	14.20	94.30	100.00	80.95	93.84	95.30	17.60	PASS
B5	99.90	100.00	100.00	100.00	99.40	01.20	99.40	100.00	100.00	100.00	97.20	04.90	**FAIL**
Day 1													
B2	93.90	100.00	80.95	93.75	96.40	17.00	93.40	100.00	80.95	93.75	96.40	19.00	PASS
B1	95.10	100.00	80.95	93.85	94.60	12.90	96.90	97.73	85.71	93.85	95.90	08.30	PASS
B5	99.80	97.92	100.00	98.55	96.80	01.30	99.00	100.00	95.24	98.55	97.20	07.40	PASS
Day 2													
B4	93.40	100.00	80.95	93.85	96.40	18.60	94.00	100.00	80.95	93.85	96.20	15.50	PASS
B2	95.60	100.00	80.95	93.75	96.40	15.00	93.80	100.00	80.95	93.75	96.40	16.00	PASS
B3	99.30	100.00	95.24	98.53	99.20	03.40	99.60	100.00	95.24	98.53	99.00	02.20	PASS
Day 3													
B6	92.90	100.00	76.19	92.06	95.00	18.40	93.30	100.00	76.19	92.06	95.30	18.60	PASS
B1	96.10	100.00	80.95	93.85	95.90	11.30	98.30	100.00	80.95	93.85	96.60	04.80	PASS
B7	99.30	100.00	90.48	97.01	98.50	03.10	99.20	100.00	90.48	97.01	98.10	04.00	PASS

Note:
CLA: Calibration accuracy
MCVA: Monte-Carlo cross validation accuracy
NCVA: Nested cross validation prediction accuracy
P: Permutation
B1, B2, B4, B5, & B7: Batch number of selected models
Days 0, 1, 2, & 3: Incubation days

Whereas TNR values for all model batches considered ranged from 71.43 to 100%, TPR values are all 100% except for a singular batch on days 3 and 1 for both brown and white eggs respectively (Tables 4.9 and 4.10). This observation inferred that the selected models are optimal in identifying fertile eggs and this is a plus for the hatchery industries who has rated fertile eggs to be of high identification importance. Although some non-fertile eggs might be misclassified as fertile eggs based on the TNR values, all fertile eggs would be mostly classified correctly in accordance to the TPR values**.** The best batch model as shown in Figure 4.10 achieved a perfect classification accuracy of 100%. Notwithstanding, this batch 5 model from the white egg set on day 0 incubation was later discovered to fail permutation testing (Table 4.10). Hence, the said model is not reliable in comparison to the median and least performing models earlier shown in Table 4.8 (see pictorial view in appendix C). This observation of a model passing CV test but failing permutation test has been reported not to be unexpected as permutation and CV verification procedures present different measures of a feature model's utility (Bijlsma et al., 2006; Westerhuis et al., 2008; Xia & Wishart, 2011). Tables 4.11 and 4.12 showed the important ratio wavelengths



True class (%)

	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	100 (48/48)	0 (0)
Actually Negative	0(0)	100 (21/21)

OVA = 100.00%

Best model, Batch 5 (3,4)

Figure 4.10: White eggs best batch model on day 0 incubation (batch 5 model 4 built with 3PCs)

selected for brown and white eggs-built models, having been optimised using the lasso modelling entering criterion (Tibshirani, 1996, 2011).

Table 4.11. Important ratio wavelengths selected for brown eggs-built classification models

Models	Important wavelength bands
Day 0	
B2	908/923, 908/927, 956/961, 1349/1368, 1642/1670
B4	1311/1627, 908, 932, 1253, 1268, 1349/1608
B7	1383/1455, 1167/1205, 1205/1220
Day 1	
B5	1220/1225, 1378/1383, 1158/1225, 1201/1225, 956/966, 1661/1666
B4	1273/1277, 1330/1340, 899/923, 1579/1613, 1579/1618, 1330/1344, 918/923, 1330/1335, 1570/1579
B7	1210/1287, 1177/1181, 1373/1378, 1301/1306, 11440/1455, 1651/1666
Day 2	
B5	1469/1488, 1249/1258, 1646/1675, 899/923
B2	1186/1201, 1493/1551, 1201/1205, 1014, 1129, 1320, 1162/1210, 1167/1201, 1167/1210, 1392/1431, 1493/1541, 1522/1551, 1522/1565, 1551/1560
B7	1239/1249, 1359/1402, 1359/1445, 1469/1512, 1148/1253, 1359/1440, 1436/1445
Day 3	
B5	1148/1244, 1527/1608, 1181/1210, 1402/1483, 1445/1483, 1531/1608, 1028, 1042, 1306/1311, 1464/1637, 1469/1608
B1	1158/1201, 1344/1474, 1359/1436, 1402/1436, 951, 1066, 1354/1474, 1153/1201, 1359/1474, 1201/1210, 1320/1436, 1325/1474, 1368/1436
B7	1416/1426, 1273/1292, 1407/1416, 1503/1575, 1598/1623

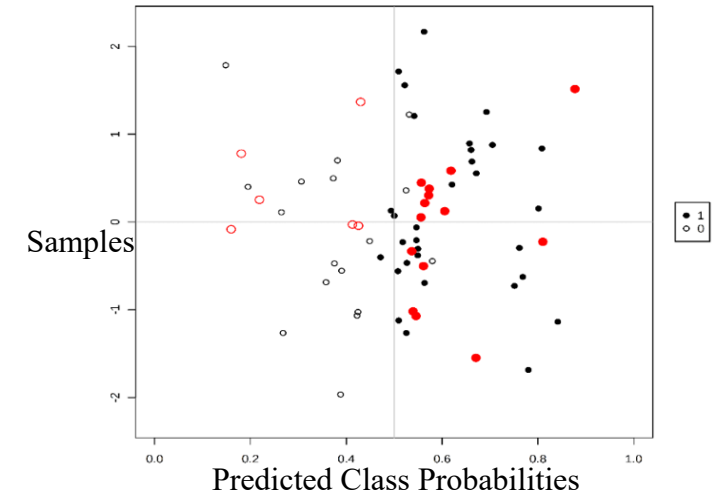
Table 4.12. Important ratio wavelengths selected for white eggs-built classification models

Models	Important wavelength bands
Day 0	
B2	1263/1277, 1613/1656, 1579/1623, 1642/1666
B1	1023/1124, 1575/1646, 1014/1129, 1306/1320, 1426/1594, 1575/1608
B5	1191/1205, 1234/1239, 1229/1234
Day 1	
B2	1316/1354, 1181/1325, 1498/1527, 1402/1565, 1450/1565
B1	1306/1311, 1388/1507, 1464/1507
B5	1541/1651, 1527/1603, 1522/1694, 1527/1570, 1527/1575, 1541/1675
Day 2	
B4	1541/1575, 1541/1608, 1407/1460, 1426/1460, 1483/1488, 1531/1541
B2	1479/1498, 1263/1268, 1268/1273, 1479/1493, 1479/1503
B3	1632/1656, 1483/1507, 951/961, 1368/1378, 1378/1416, 1560/1565, 1632/1651
Day 3	
B6	1244/1253, 1167/1191, 1181/1196, 1191/1210, 1191/1225, 1191/1292, 1253/1277, 1431/1436, 1436/1483, 1579/1594, 1632/1651, 1076, 1129, 1244, 1172/1273, 1325/1330
B1	1397/1436, 1397/1407, 1316/1325
B7	1440/1445, 1153/1158, 1158/1162, 1258/1268, 1584/1594, 1661/1666

Note: B1, B2, B4, B5, & B7: Batch number of selected models

On a final note, Figures 4.11 and 4.12 showed the verification results for both brown and white eggs respectively, using a 30% hold-out data set for testing. The hold-out data set were taken as unknown and were excluded in the calibration and validation process. In the brown egg set, while a total of 7 non-fertile (NF) and 14 fertile (F) were held out for both median and best models, 6 NF and 14 F were held out for the least model (Figure 4.11). In the white eggs set however, the number of non-fertile and fertile eggs held out were (6 NF, 13 F); (6 NF, 14 F) and (7 NF, 14 F) for the least, median and best models respectively (Figure 4.12). Wherever indicated, label 0 stands for non-fertile eggs and label 1 stands for fertile eggs.

All "hold-out" data tested on day 0 incubation period for brown eggs achieved 100% identification accuracy (Figure 4.11a to c). White eggs set on the other hand has least model achieving 100% (Figure 4.12a), median model achieving 95% (Figure 4.12b), and best model achieving 0% (Figure 4.12c) identification accuracies. For the batch 1 white eggs median model having 95% identification accuracy, there was only 1 fertile egg that was misclassified as non-fertile egg. It was further observed that the batch 5 white eggs best model, having 0% identification accuracy was the same model that had initially failed permutation testing (Table 4.10). The permutation result for the batch 5 white eggs on day 0 incubation indicates that this seemingly good calibration model shown in Figure 4.10 and verified in Figure 4.12c (CV accuracy of 97.2, OVA of 100%) could as well be resulted from random guessing and so cannot be trusted unless it can predict unknown data. Even at this, such model has instability potential of fluctuating between good prediction and wrong prediction and so should be avoided.

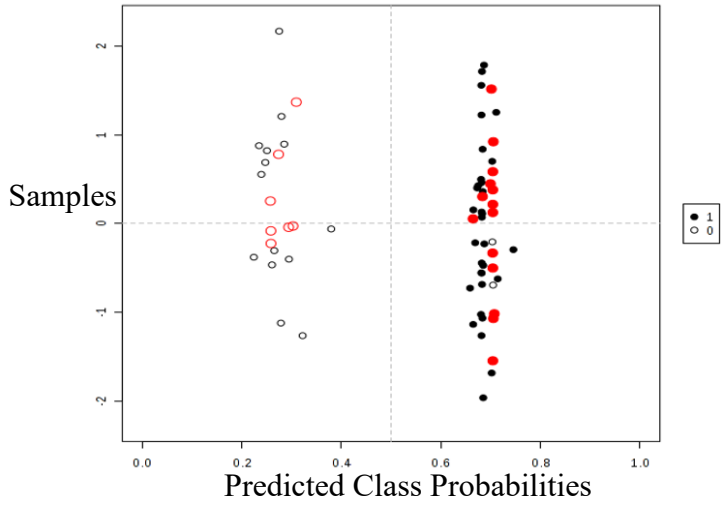


a.

Batch2(1,3)

Confusion Matrix (Hold-out)		
	0	**1**
0	6	0
1	0	14

Hold-out data prediction = 100% (20/20).

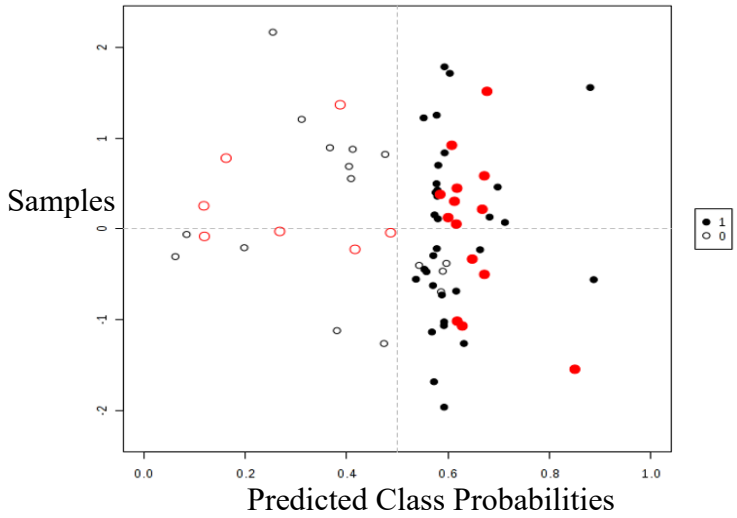


b.

Batch4(3,3)

Confusion Matrix (Hold-out)		
	0	**1**
0	7	0
1	0	14

Hold-out data prediction = 100% (21/21).



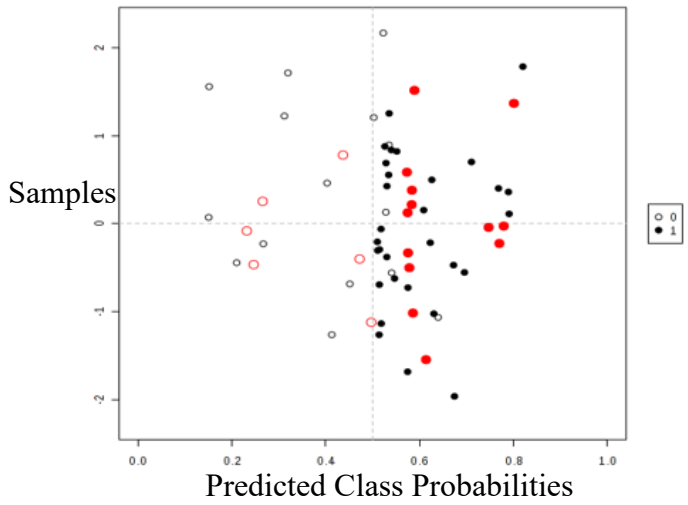
c.

Batch7(2,3)

Confusion Matrix (Hold-out)		
	0	**1**
0	7	0
1	0	14

Hold-out data prediction = 100% (21/21).

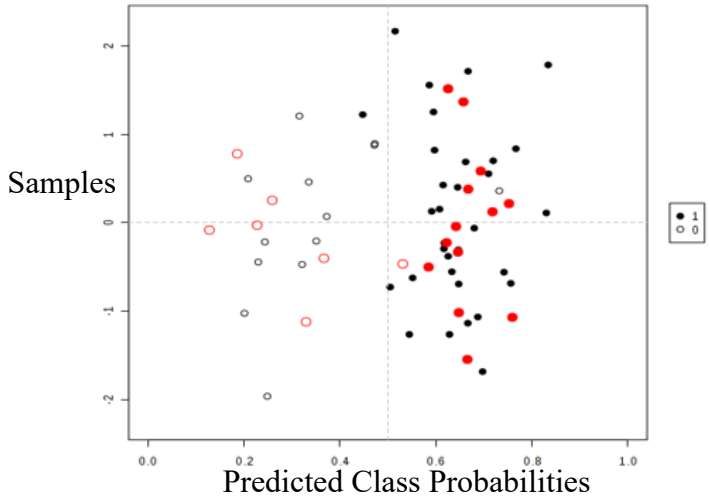
Figure 4.11: Verification models on day 0 incubation for brown eggs hold-out data set (a), least (b), median (c), best models



a.

Batch2(2,3)		
Confusion Matrix (Hold-out)		
	0	**1**
0	6	0
1	0	13

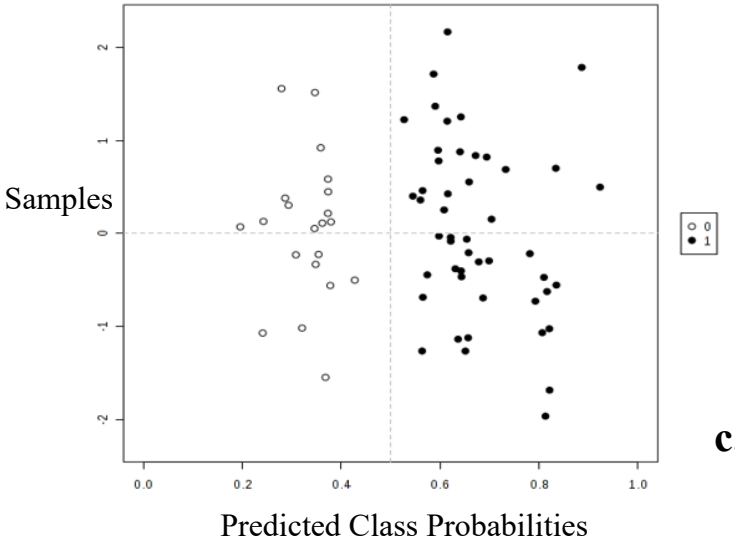
Hold-out data prediction = 100% (19/19).



b.

Batch1(2.3)		
Confusion Matrix (Hold-out)		
	0	**1**
0	6	0
1	1	13

Hold-out data prediction = 95% (19/20).



c.

Batch5(3,4)		
Confusion Matrix (Cross-validation)		
	0	**1**
0	14	0
1	0	34

100 CV accuracy = 97.20
OVA/CLA = 100 (48/48)
Hold-out data prediction = 0% (0/21).

Figure 4.12: Verification models on day 0 incubation for white eggs hold-out data set (a), least (b), median (c), best models

4.4 Conclusion

This paper has presented the details of a study carried out to investigate the appropriateness of a PLSDA learner and feature ranking technique to identify informative variables (features) and thereby selecting appropriate discriminating features from chicken egg fertility data. Due to the nature of hyperspectral imaging data in general and the specific imbalance phenomenon of chicken egg fertility data, acquired spectra data were first resampled and preprocessed prior downstream analysis. A PLSDA based learning and feature ranking algorithms were implemented within a MCCV loop to identify discriminating features for the built models. Various combination of feature subsets was learned and 5, 10, 15…100 (maximum) features were selected (at an optimal center cut-off threshold point of 0.5), as appropriate for building various discriminating models. Top 20 ratio features were also considered in the feature selection process. Model performances were evaluated from confusion matrix and area under ROC curve (AUC) computations. Models were validated using nested CV and permutation testing. Further verification was also accomplished using a 30% hold-out data set for prediction. 5PCs were considered for model building but optimal PCs eventually used for selected models do not extend beyond the first 3PCs.

The analysis results for both brown and white eggs showed that the identified features by the PLSDA algorithm are appropriate for classifying fertile eggs from non-fertile eggs during early incubation. Thus, chicken egg fertility model structure was eventually successfully developed, validated, and verified using adequate number of PCs. For all days of incubation considered and for both brown and white eggs, optima models presented were mostly rated excellent possessing AUC values in the range of 90-100%. Verification results achieving mostly 100% identification accuracy also support the above report. It was further observed that fertile eggs have higher recognition rate than non-fertile eggs. This is a great plus to the hatchery industries who has always

rated fertile eggs to be of high identification importance. Hence, the PLSDA learner-based feature selection algorithm, in conjunction with a non-parametric ROC curve analysis approach was confirmed appropriate for identifying important features towards early prediction of chicken egg fertility. Further work would entail considering other classifiers to ascertain which would best expose chicken egg fertility data structure to learning algorithms towards building an industrial online classification system.

CONNECTING TEXT TO CHAPTER 5

Upon successful identification of important features for early chicken egg fertility discrimination using PLSDA learner-based feature selection technique (chapter four), an attempt was made in chapter five to compare the performances of more classifiers with a view to determining which classifier or combination of classifiers would be well suited to exposing chicken egg fertility data structure to learning, and thereby suggesting optimal classifier(s) for translating the results of early chicken egg fertility studies to industrial practice. Understanding that the modelling approach used to identify informative variables might not be the best approach to translate the identified features into Industrial practice, up to 10 different classifiers were studied and compared in chapter 5 for adoptability potentials towards building an industrial online chicken egg fertility assessment system.

Chapter 5 would be submitted for publication to the Journal of Information Sciences as:

Adegbenjo, A. O., and Ngadi, M. (2019). Visualization of Hyperspectral Imaging (HSI) Based Chicken Egg Fertility Data for Early Discrimination Using Combination of Classifiers. *Information Sciences*

CHAPTER 5

COMPARING PERFORMANCES OF CLASSIFIER COMBINATIONS FOR ASSESSING EARLY CHICKEN EGG FERTILITY RECOGNITION USING HYPERSPECTRAL IMAGING

Abstract

In our earlier work, a PLSDA learning algorithm was used in conjunction with a non-parametric Receiver Operating Characteristic (ROC) curve analysis technique to identify and select appropriate discriminating features from chicken egg fertility data. The purpose of the present study is to finalize the optimal model structures earlier developed. Understanding that the modelling approach used to identify informative variables might not be the best approach to translate the identified features into Industrial practice, we considered in addition to PLSDA, other PLSDA algorithm variants which are: Sparse PLS discriminant analysis (sPLSDA) and orthogonal PLS discriminant analysis (OPLSDA). Some more classification algorithms like SIMCA, SVM, logistic regression (LOGREG), LDA, QDA, MDA, RF, and KNN were also tested. The outcome of our study showed that even though PLSDA algorithm was adequate in identifying informative variables in chicken egg fertility data, a simpler algorithm KNN proved more effective in exposing the chicken egg fertility data structure to learning. Whereas other classifiers like the LDA, sPLSDA, PLSDA, OPLSDA, and SVM also tied with KNN in performance for white egg discrimination, KNN being grouped among the simplest of all machine learning algorithms made it a major classifier of consideration for industrial adoptability and was thereby presented as appropriate for translating discriminative features from chicken egg fertility data into Industrial practice. Furthermore, this work presented that considering early chicken egg fertility

discrimination study as a one-class design problem, offers great potentials to learning chicken egg fertility data structure, more appropriately.

5.1 Introduction

Classification is the term used to describe division (or separation) of a group of samples into two or more classes based on characteristic features in the samples, and it has been known to be a very crucial task in pattern recognition (Swarbrick, 2012). Two basic approaches to solving classification problems in general are those of unsupervised and supervised algorithm techniques.

Notable methods used in unsupervised classification include K-means, K-medians, hierarchical cluster analysis and principal component analysis (PCA). For supervised classification, the following techniques are often employed: soft independent modelling of class analogy (SIMCA) with PCA, linear discriminant analysis (LDA), Logistic regression (LOGREG), partial least squares discriminant analysis (PLSDA) and support vector machine (SVM) classification (Swarbrick, 2012). Other unsupervised and supervised learning algorithms including but not limited to independent component analysis (ICA), fuzzy one class support vector machines (FOCSVM), random forest (RF), and associative classification have also been reported in literatures (Dong et al., 1999; Li et al., 2004; Naik & Kumar, 2011; Yu et al., 2011).

Even though it might be hard to say exactly when sparseness introduction to classifier algorithms will improve model accuracy, sparsity inclusion in algorithms have been found effective with high dimensional data structures. Sparse partial least squares discriminant analysis (sPLSDA) in particular has been noted to find useful applications where the conventional linear discriminant analysis (LDA) fails (Filzmoser et al., 2012). According to (Chung & Keles, 2010), the inclusion of sPLS into a generalized linear model produces better sensitivity in variable selection during classification in an imbalanced data scenario.

Orthogonal PLS discriminant analysis (OPLSDA) has been reported as a major development of the PLSDA method and is well known for tackling non-predictive variability in data set, thereby improving classification performance especially in cases where there exists significant difference in the within class variability between the group categories of consideration (Bylesjö et al., 2006; Trygg & Wold, 2002). Furthermore, OPLSDA was known to provide better visualization and transparency of generated model. Even though it was reported somewhere that OPLSDA would not do well on a problem where PLSDA failed (Tapp & Kemsley, 2009), earlier researchers have pointed out situations where OPLSDA might outperform PLSDA (Bylesjö et al., 2006; Mahadevan et al., 2008). It might be good to further verify these inconsistencies.

Upon completion of the important variable identification step in the modelling process, and hopefully an effective and robust set of features have been selected, the next stage in the process is generating a final stable feature model(s). This is usually done using all the data available for the feature subsets, applied to the optimal model structures (Xia et al., 2013). Optimal subset of features and model structures have earlier been identified using PLSDA algorithm in our earlier work. However, for improved performance and industrial applications, other modelling algorithms or peradventure even simpler algorithm(s) might be more appropriate.

Also, industrial research scientists might be more interested in changing analytical platforms due to various possibilities including but not limited to project completion time, algorithm complexity and even economic reasons. In the same vein, it is crucial to mention that the modelling approach used to identify informative variables might not be the best approach to translate the identified features into Industrial practice (Xia et al., 2013). In the light of the foregoing, the present study has considered in addition to using PLSDA classifier, other PLSDA algorithm variants which are: sPLSDA and OPLSDA. Some more classification algorithms like soft

independent modelling of class analogy (SIMCA), support vector machine (SVM), logistic regression (LOGREG), linear discriminant analysis (LDA) and its variants (quadratic discriminant analysis, QDA; mahalanobis discriminant analysis, MDA), random forest (RF), and K-nearest neighbours (KNN) were also tested for appropriateness in translating chicken egg fertility classification results into industrial practice.

5.2 Materials and Methods

5.2.1 Samples

A total of eight data set samples comprising of four data sets from each group of brown and white eggs respectively were visualised and analysed using the earlier listed classification algorithms. These data sets formed the least performing model structures on different days of incubation (day 0 to day 3), from our earlier work (please see our earlier work for detailed sample specifications).

5.2.2 Methodology

5.2.2.1 PCA

Principal component analysis (PCA), reported as one of the most powerful and crucial techniques in chemometrics, is well known for feature extraction, loss data compression, and dimensionality reduction (Bro & Smilde, 2014; Liu & Ngadi, 2013). PCA is likely the oldest multivariate technique with its antecedent dated back to 1901 but its modern use onset traced to Hotelling's work of 1933, in which the term principal component was first used (Hotelling, 1933; Pearson, 1901). PCA decomposes multidimensional and inter-correlated variables into a new set of orthogonal (independent) variables known as principal components (Abdi & Williams, 2010), which now serve as an important guiding results for further down stream chemical/biological analysis. Principal components are known to explain mostly the cogent information embedded in

a data set with the first component explaining the greatest amount of variability in the data, thereby possessing the potential of explaining the largest amount of crucial information. Despite this great potential and versatility of the PCA approach, it is worthy of mention here that PCA has also been reported adequate for exploratory purposes and so popularly employed as a "means to an end" tool and not necessarily an "end" in itself (Amigo et al., 2015; Shlens, 2014; Vidal & Amigo, 2012; Wold et al., 1987).

5.2.2.2 sPLSDA

The sparse version of partial least squares discriminant analysis (sPLSDA) seeks to perform dimension reduction and variable selection simultaneously (Chun & Keleş, 2010). It is well known for noise reduction capability thereby getting rid of non-informative variables from data. sPLS algorithm functions by setting noise variable contributions to zero unlike other counterpart techniques that usually set such contributions to an absolute (small) value (Filzmoser et al., 2012). The strength of sPLSDA over other similar algorithms is that of being able to perform variable selection alongside classification in a single step process. Apart from this, sPLSDA has also been reported for good computational efficiency and excellent predictive performances (Lê Cao et al., 2011). sPLSDA as implemented in this study was based on the already identified and selected ratio features and not on newly selected features. Further details of sPLSDA algorithm implementation have been described elsewhere (Calvini et al., 2015; Lê Cao,Martin, et al., 2009; Lê Cao et al., 2008).

5.2.2.3 OPLSDA

Based on the need for improvement over PLSDA multivariate method, OPLSDA was earlier developed to discriminate between two or more classes of data. A regression model is usually computed between a dependent variable Y column-vector matrix and a multidimensional X data

matrix. OPLSDA thereby employs the information embedded in the response column-vector matrix Y to decompose the X multidimensional data matrix into two different components namely the orthogonal (non-predictive) and predictive components (Bylesjö et al., 2006; Trygg & Wold, 2002; Westerhuis et al., 2010). One notable advantage of OPLSDA over PLSDA is that of the possibility of retaining only the predictive component (being the component related to the class separation) for model building, while discarding the non-predictive component (Westerhuis et al., 2010). The need for OPLS approach was propelled by the great amount of non-correlated variability existing in today's multidimensional data. The method removes systematic orthogonal variability from data when implemented singlehandedly as a preprocessing method. It additionally provides simpler models when used as an integral part of the traditional PLS modelling, with added benefit of the non-correlated variability being able to be analysed separately if need be. The versatility of OPLSDA is therefore seen in its inherent ability of handling both between class and within class variations. Furthermore, OPLS have been reported of producing more parsimonious PLS models with easy interpretation due to the clear separation of the correlated variability from the non-correlated variability (Trygg & Wold, 2002). Fuller description of the OPLS algorithm as implemented in this study can be found by consulting (Trygg, 2002; Trygg & Wold, 2002; Wold et al., 1998).

5.2.2.4 SVM

For a binary class situation, support vector machine (SVM) do seek for the maximum separating hyperplane between the two classes in question, thereby maximizing the margin between the classes' closest points (Meyer, 2017; Meyer & Wien, 2001). With the use of the sequential minimal optimization (SMO) algorithm, SVM employs the training data points (support vectors) that are closest to the optimal separating hyperplane for various higher dimensional space

projections prior to class predictions (Jason, 2016). Projections are usually controlled using kernels. Even though SVM is basically known for linear classification, it can handle non-linear separations via non-linear kernel functions integration (Ballabio & Todeschini, 2009). For further comprehension of SVM and its SMO algorithm as implemented in this study, readers can consult (Bennett & Campbell, 2000; Cristianini & Shawe-Taylor, 2000; Schölkopf et al., 1999; Smola, 2000; Vapnik & Vapnik, 1998; Weston et al., 2001).

5.2.2.5 RF

Random forest (RF) is a form of ensemble learning just like the widely known bagging and boosting of classification trees (Jason, 2016; Sun et al., 2009). RF as proposed by (Breiman, 2001), do form a collection of tree predictors (ensemble). Each tree in the ensemble gives a classification known as vote and the final tree choice in the forest is dependent on the classification having the majority votes. RF is known to be fast, user-friendly and robust against overfitting (Liaw & Wiener, 2002). Nonetheless, RF suffers from complexity limitation of the largest number of trees that can be grown with the training data and it therefore becomes practically impossible to improve its accuracy performance with unknown data (Ho, 1995). Further details about tree classifiers and their applications as adopted in this study can be obtained from (Breiman, 2001; Breiman et al., 1984; Schuermann & Doster, 1984).

5.2.2.6 LOGREG

Logistic regression (LOGREG) is a regression approach known for predicting binary response (dependent) variable, by using the maximum-likelihood ratio for computing statistical significance of variables (Hosmer & Lemeshow, 2000; Kurt et al., 2008; Özdamar, 2004). Even though LOGREG is similar to ordinary linear regression (OLR) model, it is solely appropriate for models with dichotomous dependent variable. However, the independent (explanatory) variables

under consideration could be categorical, continuous or combination of categorical and continuous variables (Hosmer & Lemeshow, 2000; Özdamar, 2004). In LOGREG analysis, it is usually assumed that all potentially important predictor variables are contained in the model and that all contained explanatory variables in the model are important (Hosmer & Lemeshow, 2000; Kleinbaum & Klein, 1994). Unlike in OLR, LOGREG does not require predictor (independent) variables to be normally distributed and it does not assume linear relationships between response and predictor variables (Subasi & Ercelebi, 2005). Further description on logistic regression and its applications can be found in (Hajmeer & Basheer, 2003; Hosmer Jr et al., 2013; Schumacher et al., 1996; Vach et al., 1996)

5.2.2.7 LDA

Linear discriminant analysis (LDA) is a supervised classification technique that produces a linear projection of m-dimensional feature instances into an n-dimensional vector space ($n < m$), such that instances of the same class become close together, while instances from different classes are wide apart from each other. LDA is the simplest of all classification methods that are based on the Baye's rule, which subsequently based on data normality and identical covariance matrices assumptions. In other words, it is generally assumed in LDA implementation that within group variabilities are structurally the same. In a situation where the structural variability assumption does not hold, linearity of the groups' separating curve fails and thereby leading to the choice of quadratic discriminant analysis (QDA) as a better alternative in such circumstance. Mahalanobis distance measure has also proved very useful when considering observation distance from one group's center to the other (Camo, 2018).

5.2.2.8 KNN

The K-nearest neighbours' (KNN) is one of the most common classification and pattern recognition approaches, adjudged as a simple but rugged learning algorithm (Chen et al., 2015; Suguna & Thanushkodi, 2010; Wu et al., 2002). The technique is used for discriminating between samples based on closest training instances in the feature learning space. An instance is classified by a majority vote of its nearest neighbours, with the sample being classified to the most common class amongst its k-nearest neighbours. For a binary classification scenario, it is a common practice choosing an odd k-value to avoid vote tie. The neighbours are usually selected from set of known correctly classified (training) samples (Shetty et al., 2010).

5.2.2.9 SIMCA

Soft independent modelling of class analogy (SIMCA), was reported as the very first discrimination concept initiated in chemistry, and remains one of the best available methods for classification purposes (Ballabio & Todeschini, 2009; Wold, 1976) in modern chemistry and related disciplines. Having understood that PCA only, might not be appropriate for class discrimination, PCA models built in this study were linked together with categorical class information to build classification models using SIMCA. SIMCA's definition of being an independent and soft class modelling technique arose from the fact that the class models are always developed separately, independent of one another, and without setting any form of hypothesis on the variable distribution. SIMCA do perform well in most applications. However, with its computations mainly based on describing within class variability over the between class variability, it usually becomes limited in finding directions for class separation (Ballabio & Todeschini, 2009) in complex cases.

5.2.3 Performance evaluation

Accuracy and performance of classifiers in this study have been evaluated from sensitivity, specificity, precision, and F-measure computations. While sensitivity (true positive rate or recall) shows the percentage of actual positives that are predicted positive, specificity (true negative rate) shows the percentage of actual negatives that are predicted negative. Precision (positive predictive value) on the other hand shows the percentage of predicted positives that are actual positives. The F-measure (F1-score) is a single metric for precision and recall, describing the harmonic mean between precision and recall. It must be noted that precision might be described in terms of the negative predictive value in a situation where the rare class has not been taken as the positive class (as it is conventionally), due to the major class being of greater recognition importance as seen to be seemingly the case in most food quality analysis. In such circumstance, recall definition would also change to favour the class of recognition importance per time, especially during computation of F1-score values, and a better informative decision can be finally taken considering average F1-score (AvF1-score). F1-score is usually computed from equation 5.1, where P is precision and R is recall. Detailed performance evaluation information for binary classification problem can be obtained from (François, 2006; He & Garcia, 2009; Sun et al., 2009).

$$F1\_score = \frac{2*P*R}{P+R} \quad \text{.....5.1}$$

5.3 Results and Discussion

Figure 5.1 showed on day 0 incubation for brown eggs, typical visualisation plots for PCA, PLSDA, OPLSDA, sPLSDA, RF, SVM, and logistic regression. Wherever indicated, label 0 stands for non-fertile eggs and label 1 stands for fertile eggs. It was observed from Figure 1 that classifiers “b” to “d” have a narrow margin of separation unlike classifiers “e” to “g” having wider

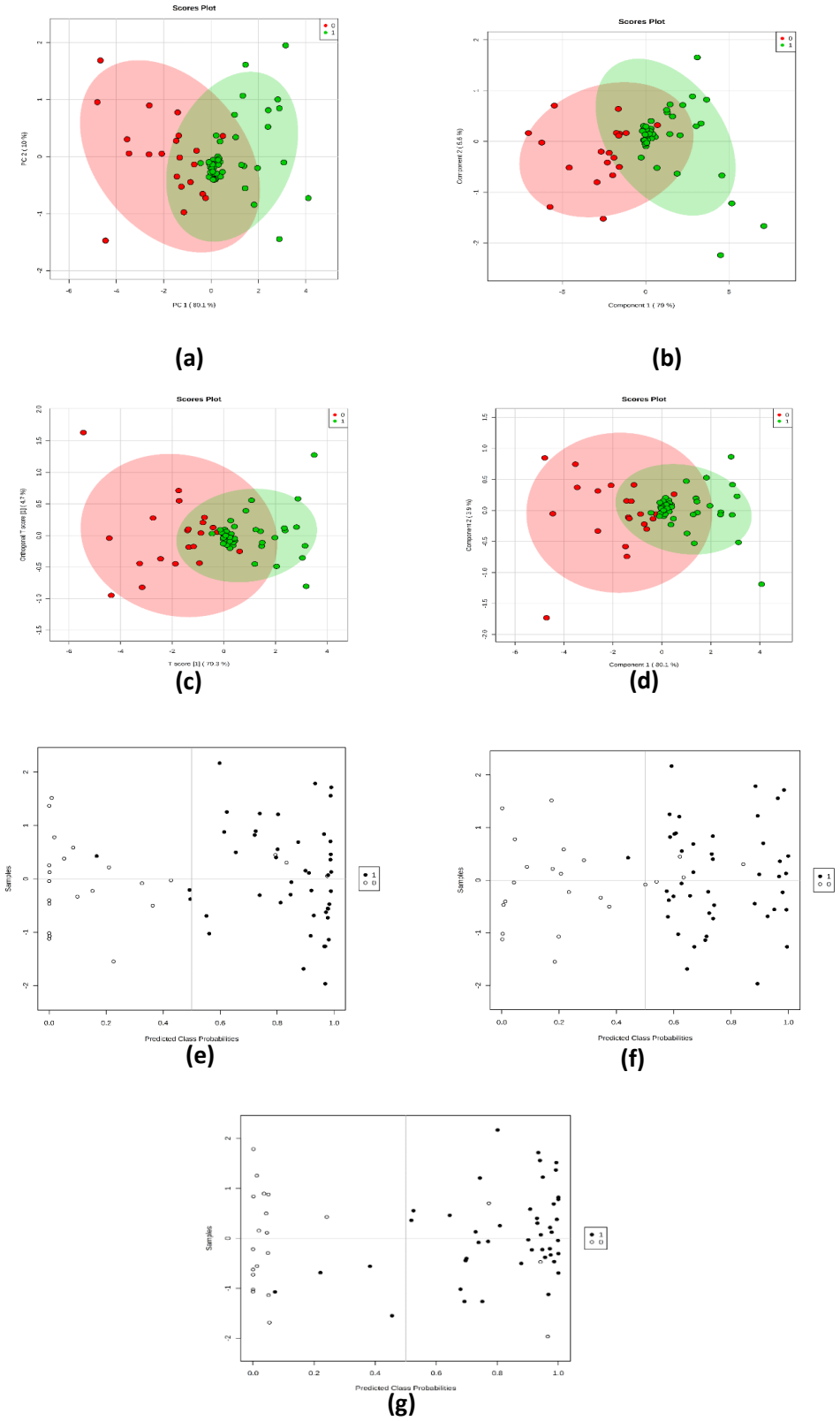


Figure 5.1. Typical visualization plots on day 0 incubation (a), PCA (b), PLSDA (c), OPLSDA (d), sPLSDA (e), RF (f), SVM (g), Logistic regression

margin of separation. Nonetheless, classifier group "b" to "d" still outperform those of "e" to "g", having more misclassification of fertile eggs (see 5.1f and 5.1g). Also, within variability were noticed in all the eggs considered with data points being wide apart, and a small portion of the fertile eggs clustering together in Figure "5.1a" to "5.1d". Table 5.1 showed the variability in percentage being explained by the first 2 PCs in the brown and white egg data sets, using PLS-based classifiers. For all days of incubation period considered, sPLSDA has the highest variability being explained in the first PC. Apart from on day 0 for brown eggs where total variability being

Table 5.1. Percentage variability in brown and white eggs from PCA and PLS-based classifiers

Brown eggs								
Inc. days	PCA		PLSDA		OPLSDA		sPLSDA	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
0	80.10	10.00	79.00	5.60	79.30	4.70	80.10	3.90
1	92.60	3.70	85.60	4.30	92.40	3.20	92.60	3.00
2	84.90	6.90	83.70	4.50	84.80	5.60	84.90	5.50
3	83.10	4.80	83.10	2.70	79.80	6.00	83.10	2.70
White eggs								
0	86.20	7.70	84.50	4.50	83.90	8.90	86.20	6.60
1	85.70	5.20	83.00	2.30	84.50	6.30	85.70	5.20
2	90.50	3.60	88.60	3.60	89.80	3.30	90.50	2.50
3	76.40	6.20	76.40	2.90	74.10	5.20	76.40	2.90

explained in the first 2 PCs with PLSDA was greater than for other classifiers, sPLSDA and OPLSDA have the greatest variability being explained in the first 2PCs for all other days of incubation for both brown and white eggs. For all total variability ties of sPLSDA and OPLSDA however, the variability being explained with sPLSDA in the first PC is always greater. In the white egg day 0 incubation for example (Table 1b), total variability explanation in the first 2PCs with both sPLSDA and OPLSDA is 92.8%. However, 86.2% of this total variability is being explained in the 1st PC with sPLSDA, whereas only 83.9% is being explained with OPLSDA in the 1st PC. Hence sPLSDA possesses potential of producing more parsimonious model than other PLS-based classifiers with chicken egg fertility data. Figure 5.2 showed sPLSDA visualization

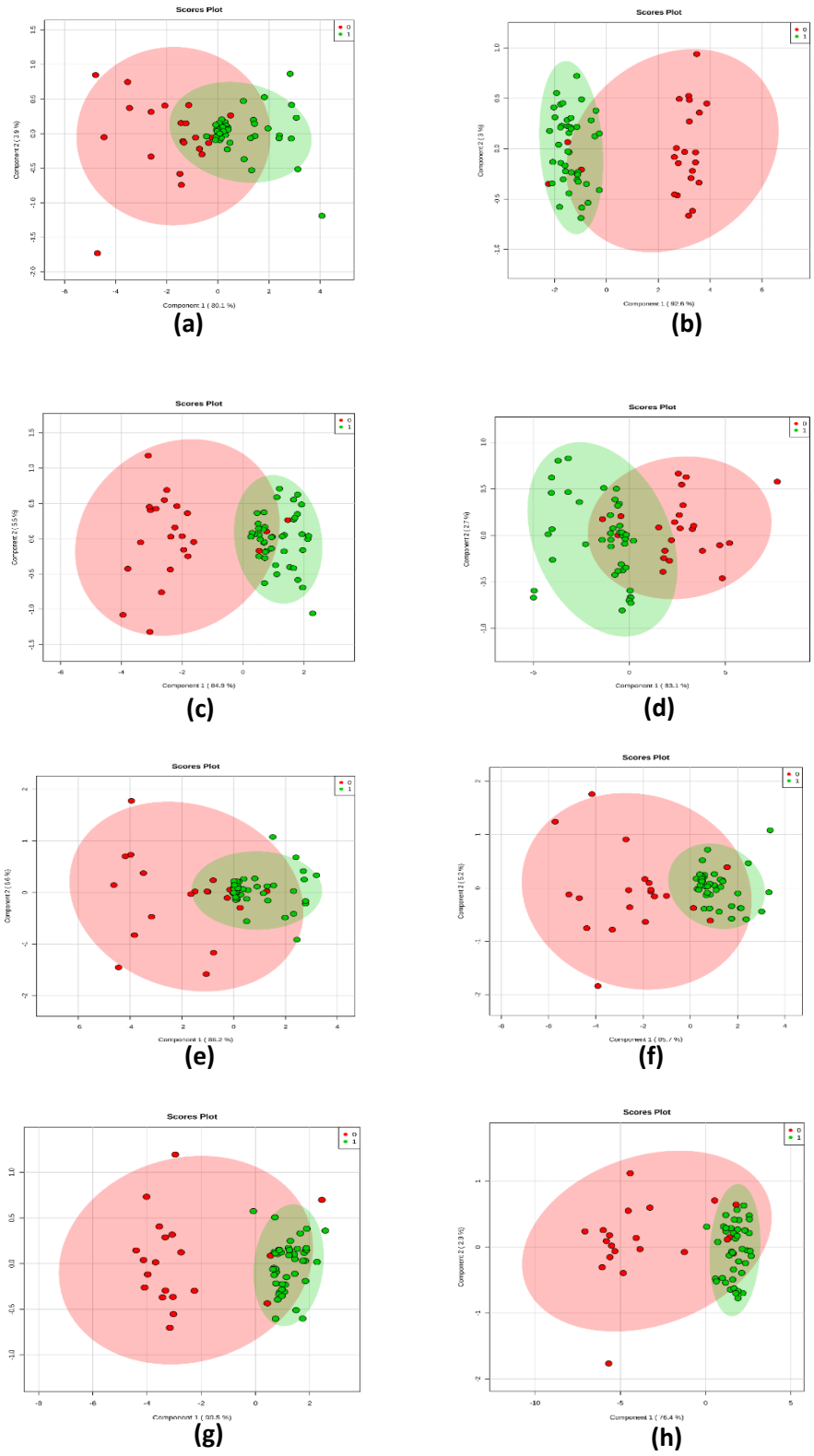


Figure 5.2. Sparse PLSDA visualization plots of brown and white eggs on different incubation periods (a), day0 brown (b), day1 brown (c), day2 brown (d), day3 brown (e), day0 white (b), day1 white (c), day2 white (d), day3 white

plots on different days of incubation for both brown and white eggs. Apart from on day 0 (Fig. 5.2 'a' and 'e'), where the separation margin was narrow as earlier mentioned, other days of incubation have wider separation margins. Also, the within variability in the eggs were reduced on other days of incubation as compared to day 0 incubation. Furthermore, the within variability with non-fertile eggs seems greater than for fertile eggs on all days of incubation.

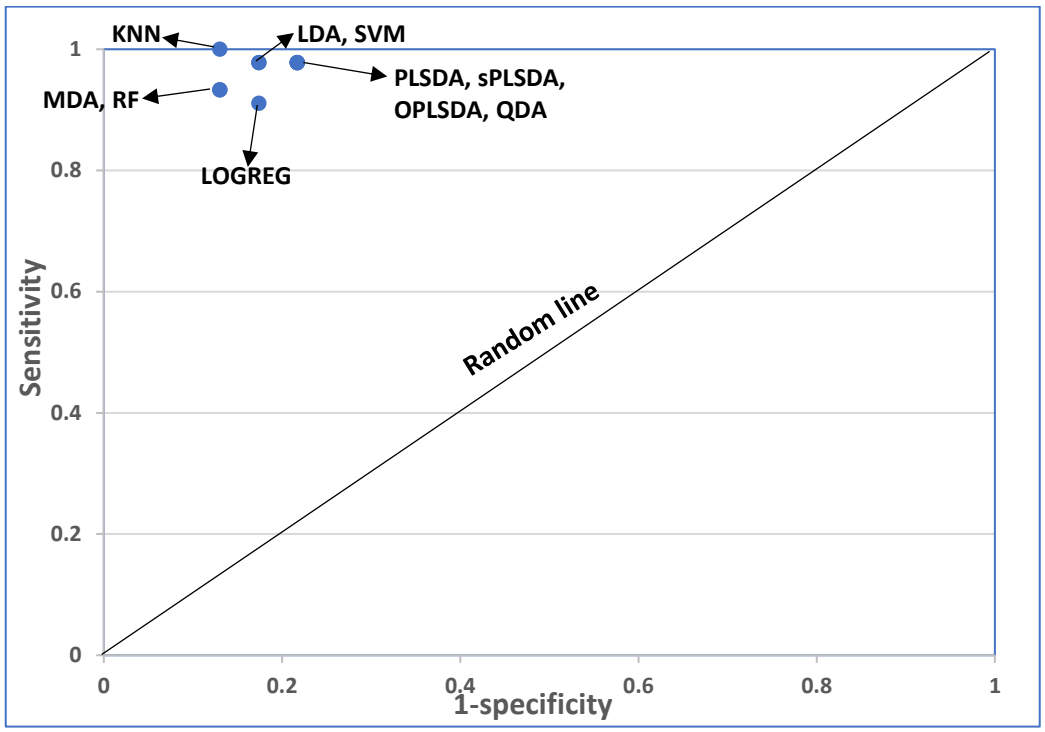
Table 5.2 showed the evaluation metric results on day 0 incubation for both brown and white eggs. For the two sets of eggs, all the PLS-based classifiers have same values for sensitivity, specificity, precision, and F1-score which are (97.80%, 78.30%, 89.80%, 93.63%) and (100.00%, 71.40%, 87.80%, 93.50%) for both brown and white eggs respectively. Despite all these classifiers

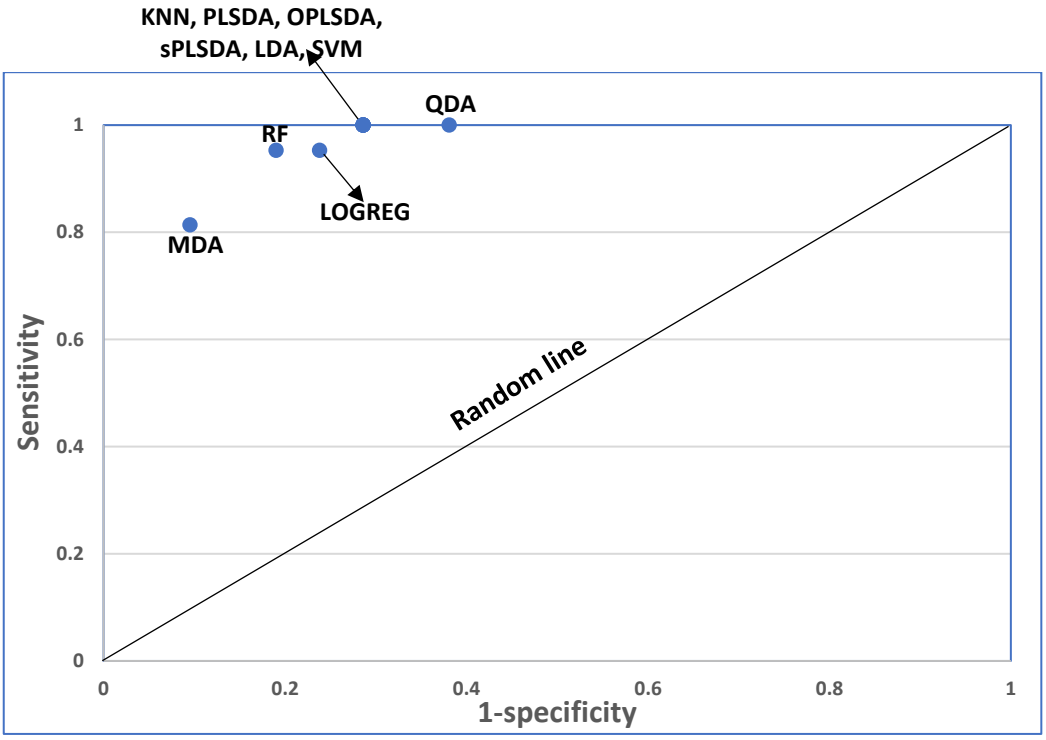
Table 5.2. Evaluation metrics (%) on day 0 incubation for brown and white eggs

Brown								
Classifier	**TP**	**TN**	**FP**	**FN**	**SEN**	**SPE**	**PRE**	**F1-score**
PLSDA	44	18	5	1	97.80	78.30	89.80	93.63
sPLSDA	44	18	5	1	97.80	78.30	89.80	93.63
OPLSDA	44	18	5	1	97.80	78.30	89.80	93.63
LDA	44	19	4	1	97.80	82.60	91.70	94.65
QDA	44	18	5	1	97.80	78.30	89.80	93.63
MDA	42	20	3	3	93.30	87.00	93.30	93.30
SVM	44	19	4	1	97.80	82.60	91.70	94.65
RF	42	20	3	3	93.30	87.00	93.30	93.30
LOGREG	41	19	4	4	91.10	82.60	91.10	91.10
KNN	45	20	3	0	100.00	87.00	93.80	96.80
White								
PLSDA	43	15	6	0	100.00	71.40	87.80	93.50
sPLSDA	43	15	6	0	100.00	71.40	87.80	93.50
OPLSDA	43	15	6	0	100.00	71.40	87.80	93.50
LDA	43	15	6	0	100.00	71.40	87.80	93.50
QDA	43	13	8	0	100.00	61.90	84.30	91.48
MDA	35	19	2	8	81.40	90.50	94.60	87.51
SVM	43	15	6	0	100.00	71.40	87.80	93.50
RF	41	17	4	2	95.30	81.00	91.10	93.15
LOGREG	41	16	5	2	95.30	76.20	89.10	92.10
KNN	43	15	6	0	100.00	71.40	87.80	93.50

perform similarly according to the shown evaluation criteria, the earlier observation in relation to variability explanation in the first 2 PCs would make sPLSDA classifier preferable among the PLS-based classifiers. For the distance-based classifiers, LDA performs better than QDA and MDA. Even though the specificity of MDA is greater than that of QDA in both sets of eggs, higher values of sensitivity in both instances made QDA a probable choice over QDA as can be buttressed in the F1-score results having higher magnitude for QDA than for MDA. For the brown eggs (Table 5.2), SVM tied with LDA in performance considering sensitivity, specificity, precision and F1-score values. Likewise, RF tied with MDA. Noting however that SVM and RF are more complex classifiers compared to LDA, QDA, and MDA would make LDA a preferable choice than SVM. Similar conclusion can be arrived at for the white eggs despite RF outperform MDA due to higher sensitivity value. Overall criteria performance from the brown eggs showed KNN as the best classifier with sensitivity, specificity, precision, and F1-score values of 100.00%, 87.00%, 93.80%, and 96.80%. Even though KNN classifier tied in performance with some other listed classifiers for the white eggs set, its being among the simplest of all machine learning algorithms made it a major classifier of consideration for industrial adoptability. For both set of brown and white eggs considered on day 0 incubation, MDA, RF, and logistic regression were the least performing classifiers and might be excluded from consideration for industrial adoptability.

The explanation above was further buttressed in the ROC of classifiers Figure 5.3, in which all classifiers were above the random line, and so none of the classifier accuracies was because of random guess. Notwithstanding, KNN has highest accuracy being positioned topmost in the ROC space for brown eggs (Figure 4a). Other top classifiers include LDA, SVM, PLS based classifiers, and QDA. Least performing classifiers positioned just below the top classifiers were logistic regression, MDA, and RF. Similar observation was seen in the white egg ROC space Figure 4b.





(b)

Figure 5.3. ROC of classifiers on day 0 incubation (a), brown eggs (b), white eggs

MDA performed the least being positioned below all other classifiers. Next in performance were LOGREG and RF. Top performing classifiers include KNN, PLS based classifiers, LDA, SVM, and QDA. Even though QDA was positioned on the same top sensitivity point with the KNN group, it was positioned farther on the right and so would perform less optimally than the KNN group classifiers on its left side. From the KNN group of classifiers, SVM might be excluded being a complex algorithm than others. sPLSDA as earlier said would be preferable than its counterparts on parsimony consideration, and KNN being among the simplest of all available classifiers (with little or no training time) would be preferred above LDA.

Figure 5.4 showed the F-measure evaluation metric curves. While Figures 5.4 a, b, and e showed individual F1-score chart for fertile and non-fertile egg recognition, together with their average for brown eggs, Figures 5.4 c, d, and f showed similar F1-score charts for white eggs. It was observed for brown eggs (Figures 5.4 a, b, and e), that KNN classifier performed optimally on all incubation periods of days 0 to 3, and LOGREG performed least. Following LOGREG in least performance was MDA on all other days of incubation apart from little deviation on day 0 incubation, in which it outperformed PLSDA, OPLSDA, and QDA but performed equally as RF with AvF1-score of 90.15% (Figure 5.4e). Other optimal performing classifiers after KNN on day 0 incubation for brown eggs based on AvF1-score values include sPLSDA, LDA, and SVM. While sPLSDA stood out again as best classifier among the PLS-based classifiers, LDA would be preferred above SVM due to its simplicity despite they have equal performance with AvF1-score of 91.51%. Figure 5.4e also showed that despite all other classifiers apart from LOGREG and MDA performed equally as KNN (94.84% AvF1-score) on incubation days 1 to 3, they underperformed on day 0 incubation and thereby retaining KNN classifier as the best choice for industrial application consideration.

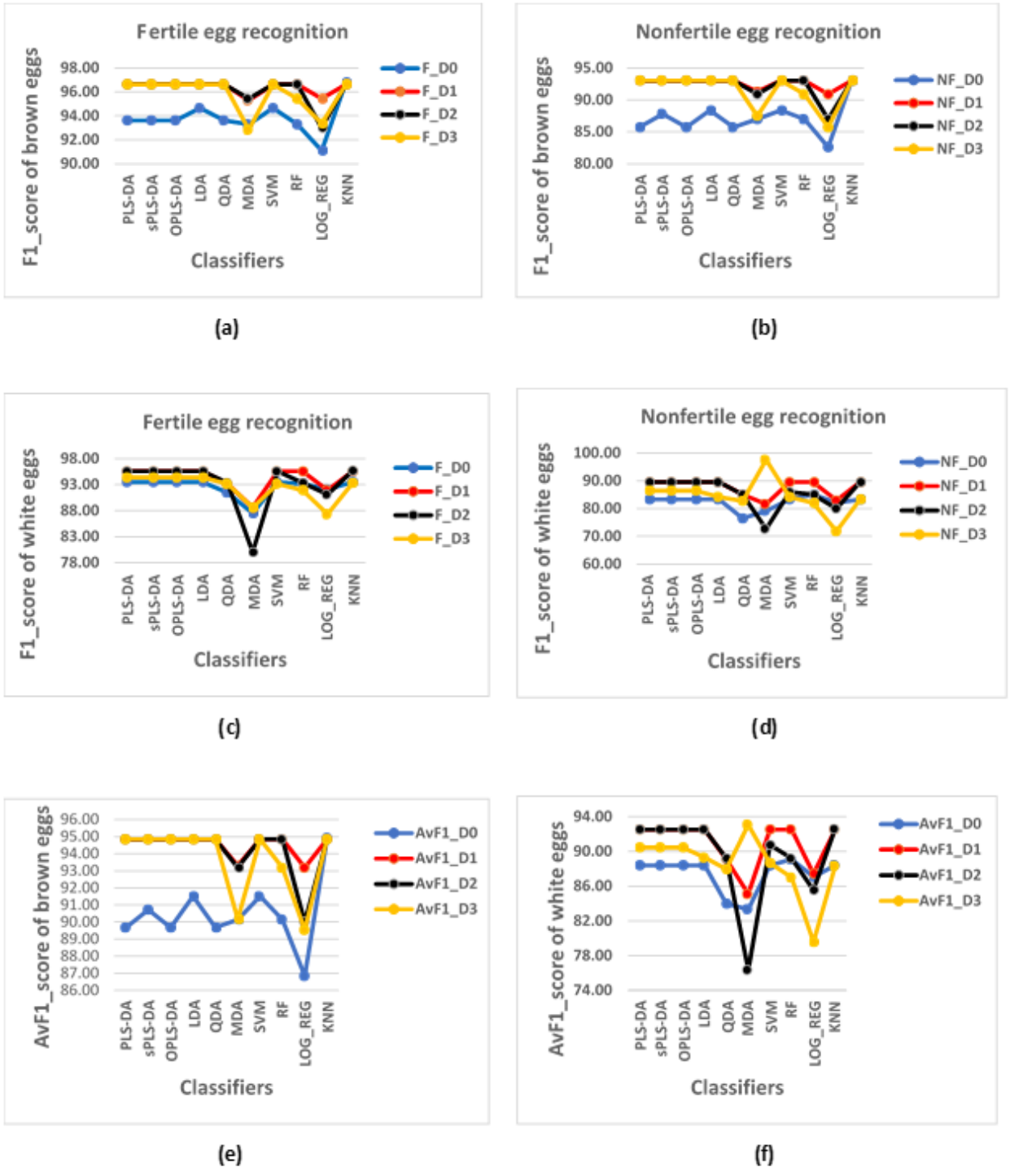


Figure 5.4: Evaluation metric curves (a), F1-score for brown fertile eggs (b), F1-score for brown non-fertile eggs (c), F1-score for white fertile eggs (d), F1-score for white non-fertile eggs (e), Average F1-score for brown eggs (f), Average F1-score for white eggs

From Figure 5.4c for fertile egg recognition, F1-score values for white eggs showed similar results as with brown eggs for days 0 to 2 incubations by having KNN on the top list. Even though sharing performance with some other classifiers on days 0 and 1, KNN singly top the list on day 2 incubation and was next after some classifiers on day 3 incubation. For non-fertile egg recognition (Figure 5.4d), RF classifier top the list with KNN coming after with some other classifiers on day 0 incubation. KNN with some other classifiers were optimal on days 0 and 1 incubations but MDA became optimal on day 3 incubation. Thus, the classifiers' performance was not consistent for both fertile and non-fertile egg recognition for white eggs. This unstable classifier performance was reflected in the AvF1-score in Figure 5.4f, in which RF was the optimal classifier on day 0 incubation, being followed by KNN with some other classifiers. While KNN was optimal with some other classifiers on day 1 incubation, it was the singly optimal classifier on day 2 and MDA became optimal on day 3 incubation. These classifier performances for white eggs seemed related to the considered white egg data being noisier than the brown eggs. It might also be related to the greater within variability existing in the white non-fertile egg set as was earlier visualised from Figure 5.2 'e' to 'h'. Nevertheless, it was clear from previous day 0 brown and white egg analysis (Table 5.2; Figure 5.3), and present white egg analysis (Figure 5.4 c, d, and f) that the classifiers considered were much better exposed to learning the fertile egg data structure than the non-fertile eggs data structure (precision and recall computations discussed for the earlier said analysis were mostly based on fertile eggs taken as being of greater recognition importance than the non-fertile eggs, and this is consistent with the present industrial requirements).

Figures 5.5 and 5.6 showed the global PCA scores and loadings plots towards a SIMCA classification analyses. While the scores plot for each day of incubation showed the clustering patterns in the egg data, the loading plots showed the ratio features responsible for the observed

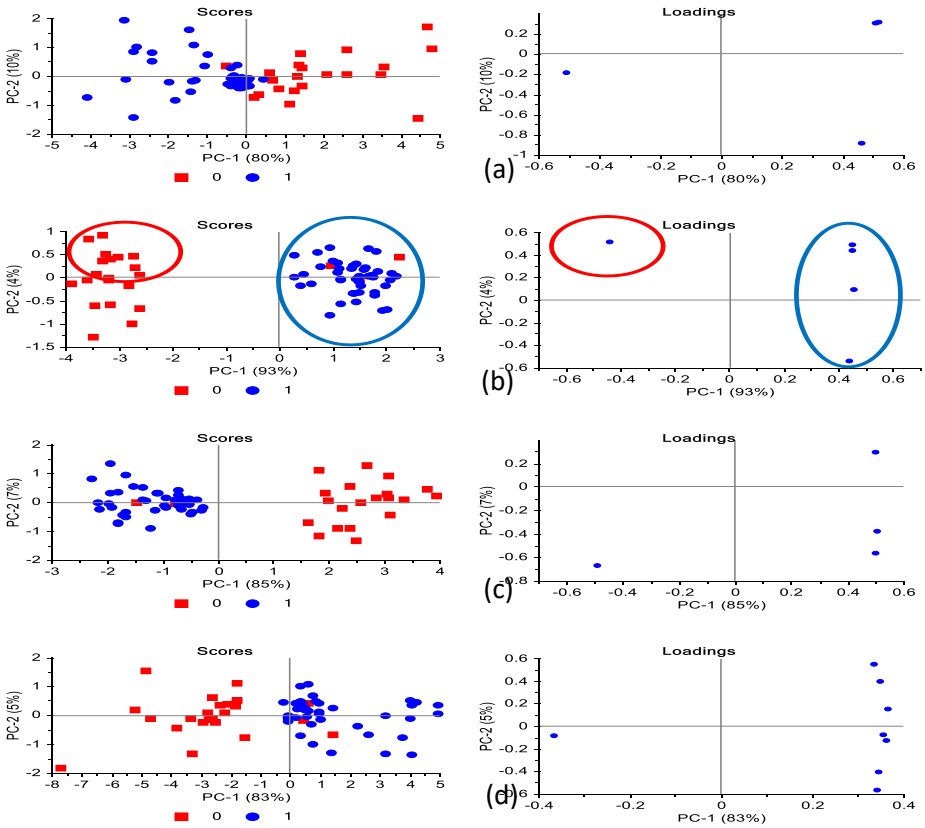


Figure 5.5. PCA scores and loadings plots for brown eggs (a), day 0 (b), day1 (c), day 2 (d), day 3

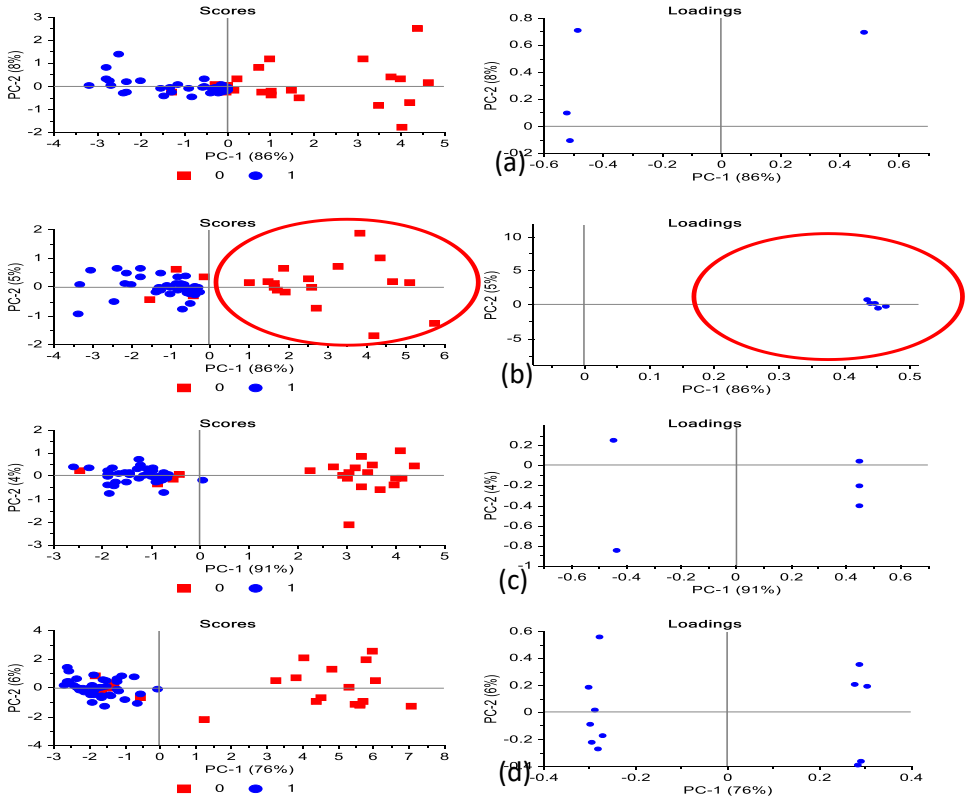


Figure 5.6. PCA scores and loadings plots for white eggs (a), day 0 (b), day1 (c), day 2 (d), day 3

clustering. As earlier mentioned with some classifiers above, the global PCA plots here also showed separation margin being narrower on day 0 incubation (Figures 5.5a and 5.6a), but wider on all other incubation days, for both brown and white eggs (Figures 5.5b-d and 5.6b-d). Figure 5.5b showed that while four ratio features were responsible for the clustering in the fertile eggs, only one ratio feature was responsible for the clustering in non-fertile eggs on this day 1 incubation. Whereas 3, 3, and 1 features respectively were responsible for non-fertile eggs clustering on days 0, 2, and 3 incubations for brown eggs (Figure 5.5 a, c, d), the clustering in fertile eggs on the same days of incubation were due to 1, 1, and 7 features. It was observed from Figure 5.6b that all the five features selected were responsible for non-fertile eggs clustering and none was responsible for fertile egg clustering. Understanding that not having enough variables to explain variability in specific class(es) of data might impede classification accuracy, more variable features were considered based on their degree of importance in learning and recognising respective classes of eggs under consideration. Nonetheless, classification potential seems not to improve with this optimization step as can be seen in the model optimization testing Figure 5.7, for incubation day 1 brown eggs. It was clear from all variable optimization situations considered that there was no improvement in classification potential. There were 3 non-fertile eggs being incorrectly clustered with the fertile eggs in all four cases (Figures 5.7 a-d). Figure 5.7c even has one fertile egg being clustered into the non-fertile eggs' quadrant.

Variables of high importance for recognising fertile eggs were first increased from 4 (Figure 5.5b loadings) to 10 (Figure 5.7a loadings) and retaining only one variable of high importance for recognising the non-fertile eggs. Despite the increased fertile egg description variables did not improve good classification potential, it furthered impede model's parsimonious performance, as can be seen in the variance being explained in the 1st PC decreasing from 93% in Figure 5.5b to

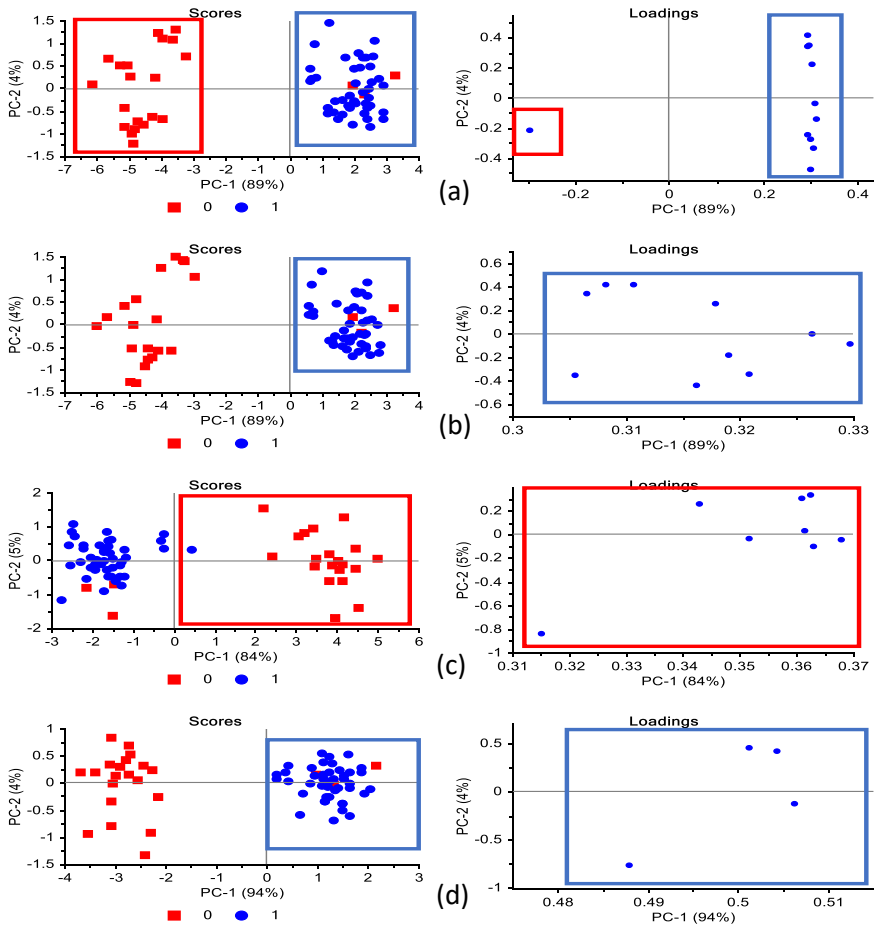


Figure 5.7. Model optimization testing for day 1 brown eggs (a), 10Fvar vs 1NFvar (b), 10Fvar (c), 8NFvar (d), optimal 4Fvar vs 0NFvar

89% in Figure 5.7a. When the one non-fertile egg description variable was excluded but retaining all 10 fertile egg description variables (Figure 5.7b), it was observed that there was no change in performance (clustering pattern remains relatively the same as well as percentage variability explanation in the 1st PC being still 89%), and hence leading to the inference that absence of non-fertile egg descriptive variable(s) in a model, might have no noticeable effect on classification performance and model's parsimony (Figure 5.7b). To buttress this fact, all fertile egg descriptive variables were removed from the model and replaced with 8 variables of high importance for non-fertile egg description (Figure 5.7c). It was observed that classification performance potential begins to degrade as one fertile egg is now entering the non-fertile eggs quadrant. Furthermore,

parsimonious performance was degraded too as percentage variability explanation in the 1st PC decreased from 89% to 84%. It was therefore concluded that while the inclusion of non-fertile egg description variables might be optional in chicken egg fertility classification modelling, inclusion of appropriate number of fertile egg description variables in the model is very critical.

We therefore lastly tested the inclusion of only top 4 variables of high importance for fertile egg description. The global PCA model obtained in Figure 5.7d showed clustering performance improvement as compared to the previous in Figure 5.7c and showed parsimonious performance improvement as can be seen from the variability explanation percentage in the 1st PC being increased from 84% (Figure 5.7c) to 94% (Figure 5.7d). This performance is also better than that reported in Figure 5.5b (PC-1, 93%) when one non-fertile egg description variable was included in the model with the 4 fertile egg description variables. The four-identified fertile egg description variables as depicted in Figure 5.7d are ratio features in the NIR wavelength bands b64/b69, b55/b69, b68/b69, and b160/161. These are corresponding to NIR ratio wavelengths (nm): 1201/1225, 1156/1225, 1220/1225, and 1661/1666.

Figure 5.8 showed typical projected PCA bi-plots of brown eggs in their respective classes, and on different days of incubation. This combined scores and loading plots showed clearly the within variability existing in each category of eggs, together with the variables responsible for such variability. For example, in Figure 5.8a, three ratio features b3/b6, b3/b7, and b95/b99 were responsible for the fertile egg clustering in the top and bottom right-hand quadrants of the plot. Likewise, feature b156/b162 was responsible for the fertile egg clustering pattern mostly in the bottom left-hand quadrant. For the non-fertile eggs (Figure 5.8b), b156/b162 was responsible for the pattern mostly at the top left-hand corner of the plot. Features b3/b6, b3/b7, and b95/b99 on the other hand were responsible for the non-fertile egg clustering pattern in the top and bottom

right-hand quadrants. It was further observed that the amount of within variability being explained in the 1st PC for non-fertile eggs are greater than that being explained for fertile eggs on all days of incubation, except on day 3 (Figure 5.8a-f). This greater within variability existing in the non-fertile eggs more than in the fertile eggs might be responsible for degraded classifier performance in recognising accurately the non-fertile eggs.

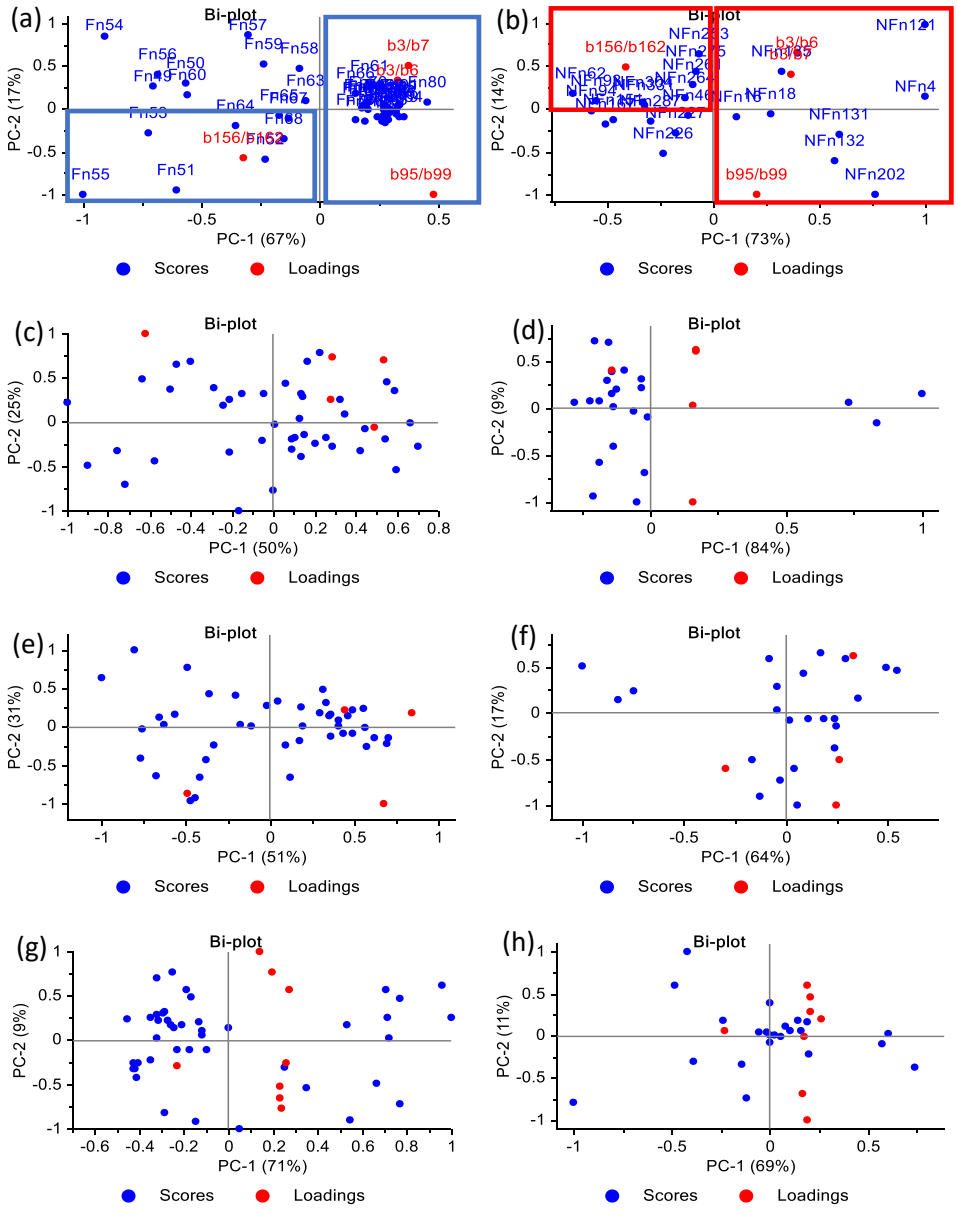


Figure 5.8. Typical projected PCA bi-plots for brown eggs (a), day 0 fertile (b), day 0 non-fertile (c), day 1 fertile (d), day1 non-fertile (e), day 2 fertile (f), day 2 non-fertile (g), day 3 fertile (h), day3 non-fertile

Tables 5.3 and 5.4 shows SIMCA classification results for both brown and white eggs respectively, on days 0 and 1 incubation periods. SIMCA classification results for other days of incubation were as shown in appendix D. From Table 5.3 on day 0 incubation, 3 samples (F3, F6, and F7) were not recognised by any of the modelled classes (M_F and M_NF), 15 samples (NF1, NF8-NF11, NF14, NF17, F1, F2, F4, F5, F8, F12, F15, and F27) belonged to one class and 50 samples (NF2-NF7, NF12, NF13, NF15, NF16, NF18-NF23, F9-F11, F13, F14, and F16-F45) belonged to more than one class. On day 1 (Table 5.3) incubation however, only one sample F9 was not recognised as being a member of any of the modelled classes. Whereas 21 samples (NF1-NF14, NF17-NF23) belonged solely to one modelled class (M_NF1), 43 samples (NF15, NF16, F10-F29, and F31-F43) were membership of more than one modelled class, and F30 belonged to modelled class M_F1. Similarly, for white eggs (Table 5.4) on day 0 incubation, 2 samples (F1, and F6) belonged to none of the modelled classes, 18 samples belonged to one class, and 44 samples belonged to more than one class. In the same vein for day 1 incubation for white eggs, 2 samples F4 and F6 showed no membership to any of the modelled classes. Whereas 19 samples showed membership of one modelled class, 43 samples showed membership of both modelled classes M_F1 and M_NF1.

Observation from Tables 5.3 and 5.4 (especially from day 1 incubation) showed there is little or no descriptive information inside non-fertile eggs to aid effective classification modelling. Efforts should therefore be concentrated on learning the fertile eggs which potentially carry distinguishable information. Hence, it was inferred that chicken egg fertility modelling might benefit from a one-class (single model projection) problem design, in which predictive model is usually developed for a single class, because either the other class(es) is too costly to acquire or very difficult to learn (Feng & Chen, 2008; Khan & Madden, 2010). If an unknown sample

therefore belongs to the modelled class, it is accepted, and if otherwise, such sample is rejected.

Table 5.3. Typical SIMCA classification table for brown eggs on different days of incubation

Day 0							Day 1					
D0Br	**M_F0**	**M_NF0**	**D0Br**	**M_F0**	**M_NF0**		**D1Br**	**M_F1**	**M_NF1**	**D1Br**	**M_F1**	**M_NF1**
NF1		●	F12	●			NF1		●	F12	●	●
NF2	●	●	F13	●	●		NF2		●	F13	●	●
NF3	●	●	F14	●	●		NF3		●	F14	●	●
NF4	●	●	F15	●			NF4		●	F15	●	●
NF5	●	●	F16	●	●		NF5		●	F16	●	●
NF6	●	●	F17	●	●		NF6		●	F17	●	●
NF7	●	●	F18	●	●		NF7		●	F18	●	●
NF8		●	F19	●	●		NF8		●	F19	●	●
NF9		●	F20	●	●		NF9		●	F20	●	●
NF10		●	F21	●	●		NF10		●	F21	●	●
NF11		●	F22	●	●		NF11		●	F22	●	●
NF12	●	●	F23	●	●		NF12		●	F23	●	●
NF13	●	●	F24	●	●		NF13		●	F24	●	●
NF14		●	F25	●	●		NF14		●	F25	●	●
NF15	●	●	F26	●	●		NF15	●	●	F26	●	●
NF16	●	●	F27	●	●		NF16	●	●	F27	●	●
NF17		●	F28	●	●		NF17		●	F28	●	●
NF18	●	●	F29	●	●		NF18		●	F29	●	●
NF19	●	●	F30	●	●		NF19		●	F30	●	
NF20	●	●	F31	●	●		NF20		●	F31	●	●
NF21	●	●	F32	●	●		NF21		●	F32	●	●
NF22	●	●	F33	●	●		NF22		●	F33	●	●
NF23	●	●	F34	●	●		NF23		●	F34	●	●
F1	●		F35	●	●		F1	●	●	F35	●	●
F2	●		F36	●	●		F2	●	●	F36	●	●
F3			F37	●	●		F3	●	●	F37	●	●
F4	●		F38	●	●		F4	●	●	F38	●	●
F5	●		F39	●	●		F5	●	●	F39	●	●
F6			F40	●	●		F6	●	●	F40	●	●
F7			F41	●	●		F7	●	●	F41	●	●
F8	●		F42	●	●		F8	●	●	F42	●	●
F9	●	●	F43	●	●		F9			F43	●	●
F10	●	●	F44	●	●		F10	●	●			
F11	●	●	F45	●	●		F11	●	●			

Table 5.4. Typical SIMCA classification table for white eggs on different days of incubation

Day 0							Day 1					
D0Wh	**M_F0**	**M_NF0**	**D0Wh**	**M_F0**	**M_NF0**		**D1Wh**	**M_F1**	**M_NF1**	**D1Br**	**M_F1**	**M_NF1**
NF1	•	•	F12	•	•		NF1		•	F12	•	•
NF2		•	F13	•	•		NF2		•	F13	•	•
NF3		•	F14	•	•		NF3	•	•	F14	•	•
NF4		•	F15	•	•		NF4	•	•	F15	•	•
NF5	•	•	F16	•	•		NF5	•	•	F16	•	•
NF6	•	•	F17	•	•		NF6	•	•	F17	•	•
NF7	•	•	F18	•	•		NF7	•	•	F18	•	•
NF8	•	•	F19	•	•		NF8	•	•	F19	•	•
NF9	•	•	F20	•	•		NF9		•	F20	•	•
NF10		•	F21	•	•		NF10		•	F21	•	•
NF11		•	F22	•	•		NF11		•	F22	•	•
NF12		•	F23	•	•		NF12		•	F23	•	•
NF13		•	F24	•	•		NF13		•	F24	•	•
NF14		•	F25	•	•		NF14		•	F25	•	•
NF15		•	F26	•	•		NF15		•	F26	•	•
NF16		•	F27	•	•		NF16		•	F27	•	•
NF17	•	•	F28	•	•		NF17		•	F28	•	•
NF18	•	•	F29	•	•		NF18		•	F29	•	•
NF19	•	•	F30	•	•		NF19		•	F30	•	•
NF20		•	F31	•	•		NF20		•	F31	•	•
NF21		•	F32	•	•		NF21	•	•	F32	•	•
F1			F33	•	•		F1	•	•	F33	•	•
F2	•		F34	•	•		F2		•	F34	•	•
F3	•		F35	•	•		F3	•		F35	•	•
F4	•	•	F36	•	•		F4			F36	•	•
F5	•		F37	•	•		F5	•		F37	•	•
F6			F38	•	•		F6			F38	•	•
F7	•		F39	•	•		F7	•		F39	•	•
F8	•		F40	•	•		F8		•	F40	•	•
F9	•	•	F41	•	•		F9	•	•	F41	•	•
F10	•		F42	•	•		F10	•	•	F42	•	•
F11	•	•	F43	•	•		F11	•	•	F43	•	•

Table 5.5 (which must be studied alongside Tables 5.3 and 5.4 for better understanding), showed SIMCA classification results, when chicken egg fertility data analysis was handled as a one-class design problem. It was observed from Table 5.5 a and b that unknown samples projection to fertile egg models gave maximum sensitivity value of 100% apart from white egg data on day

1 incubation that gave 95.10% sensitivity. Projection to non-fertile egg models on the other hand, gave very low sensitivities ranging from 0.00% to 22.00%. Even though, projection to non-fertile egg models gave maximum specificity values of 100% all through, corresponding sensitivity values as earlier stated are very low as compared to corresponding specificity values (ranging from 30.40% to 91.30%), should the new samples be projected to fertile egg models.

Table 5.5 SIMCA percentage one-class classifier performance, cross validation results for brown and white eggs

(a)								
Model (N)	**TP**	**TN**	**FP**	**FN**	**UNC**	**SEN**	**SPE**	**PRE**
M_F0 (68)	42	7	16	**0**	**3**	100.00	30.40	72.40
M_NF0 (68)	**8**	23	0	35	**2**	18.60	100.00	100.00
M_F1(66)	42	21	2	**0**	**1**	100.00	91.30	95.50
M_NF1 (66)	**1**	23	0	41	**1**	2.40	100.00	100.00
M_F2 (66)	42	20	3	**0**	**1**	100.00	87.00	93.30
M_NF2 (66)	**4**	23	0	38	**1**	9.80	100.00	100.00
M_F3 (66)	41	7	16	**0**	**2**	100.00	30.40	71.90
M_NF3 (66)	**9**	23	0	32	**2**	22.00	100.00	100.00
(b)								
M_F0 (64)	41	12	9	**0**	**2**	100.00	57.10	82.00
M_NF0 (64)	**6**	21	0	35	**2**	14.60	100.00	100.00
M_F1 (64)	39	14	7	**2**	**2**	95.10	66.70	84.80
M_NF1 (64)	**3**	21	0	38	**2**	7.30	100.00	100.00
M_F2 (65)	44	17	4	0	-	100.00	81.00	91.70
M_NF2 (65)	0	21	0	44	-	0.00	100.00	0.00
M_F3 (63)	42	16	5	0	-	100.00	76.20	89.40
M_NF3 (63)	0	21	0	42	-	0.00	100.00	0.00

Although precision values for non-fertile egg projected samples achieved up to 100% in most instances, the very low sensitivity values pose a disadvantage in considering such models for an industrial application, as most fertile eggs (which is of uttermost recognition importance in the hatchery industries) would be majorly misclassified. The egg samples in bold red are all additively related to the unclassified samples in bold blue font. For example, all 3 supposedly false negatives in the brown eggs, for the M_F0 (Table 5.5a) were really, unclassified samples (belonging to

neither of the modelled classes). In the same vein, there were supposedly 10 true positives for the M-NF0, but 2 of them were unclassified samples (consider in conjunction with table 5.3). Whereas, there were 2 unclassed samples in all of incubation days 0 and 1 data cases for white eggs (Table 5.5b), days 2 and 3 incubation data have no unclassified samples. The unclassified samples are usually removed during model optimization and they were therefore excluded from sensitivity, specificity and precision computations.

It was evident from the foregoing that handling early chicken egg fertility detection study as a one-class design problem has very promising potentials, using fertile egg class as the learning/training class. Some more specific one-class modelling algorithms including but not limited to the one-class support vector machine (OCSVM), fuzzy one-class support vector machines (FOCSVM), autoencoders, single-class mini-max probability machine (SCMPM), and one-class KNN (Chawla et al., 2004; Ghaoui et al., 2003; Munroe & Madden, 2005; Schölkopf et al., 2001; Yu et al., 2011) might offer additional learning advantages.

5.4 Conclusion

This paper has presented the details of a study carried out to compare the performances of eleven different classification algorithms in a way to determine which classifier or combination of classifiers might be well suited for translating the research outcome of early chicken egg fertility predictive modelling into industrial practice. Classification algorithms including PLSDA, sPLSDA, OPLSDA, LDA, QDA, MDA, SVM, RF, LOGREG, and KNN were evaluated from sensitivity, specificity, precision, F-measure, and ROC computations. SIMCA analysis, though discussed separately because of its different implementation structure, did not outperform any other classifier on day 0 incubation in its present form. It was observed that all classifiers evaluated for day 0 incubation, were all positioned above the random line in the ROC plots for both brown

and white eggs. Hence, the accuracy of the classifiers was not because of a random guess, but all classifiers performed above average. For both brown and white eggs, all PLS-based classifiers performed similarly having same sensitivity, specificity, precision, and F1-score values. However, from parsimonious consideration point of view, sPLSDA was recommended preferable among the PLS-based classifiers due to its ability to explain largest amount of variability in the first 2PCs. LDA performed better among the distance-based classifiers. Even though SVM and RF tied at some instances with distance-based classifiers, LDA was still found preferable being a simpler classifier than SVM and RF.

Overall criteria performance from the brown eggs on day 0 incubation showed KNN as the best classifier with sensitivity, specificity, precision, and F1-score values of 100.00%, 87.00%, 93.80%, and 96.80%. Even though KNN classifier tied in performance with some other listed classifiers for the white egg set, its being among the simplest of all machine learning algorithms made it a major classifier of consideration for industrial adoptability. For both set of brown and white eggs considered on day 0 incubation, MDA, RF, and LOGREG were the least performing classifiers and might be excluded from consideration for industrial adoptability. Despite all other classifiers apart from LOGREG and MDA performed equally as KNN (94.84% AvF1-score) on incubation days 1 to 3 for brown eggs, their underperformance on day 0 incubation made KNN classifier the best choice for industrial application consideration. Even though there were some instability in classifier performance for white eggs set (possibly related to noise and greater within variability existing in the non-fertile white eggs), it was generally clear from our analyses that all the classifiers considered were much better exposed to learning the fertile egg data structure than the non-fertile eggs data structure. This observation was also consistent with the outcome of SIMCA analysis especially from day 1 incubation onwards.

Results output from SIMCA analysis further identified four variables of high importance for fertile eggs recognition which are ratio features in the NIR wavelength bands b64/b69, b55/b69, b68/b69, and b160/161 (corresponding to NIR ratio wavelengths (nm): 1201/1225, 1156/1225, 1220/1225, and 1661/1666). It was eventually ascertained that there is little or no descriptive information inside non-fertile eggs to aid effective classification modelling, and so efforts should be concentrated on learning the fertile eggs, carrying potential distinguishable information. It was therefore proposed that chicken egg fertility modelling might benefit from a one-class (single model projection) problem design, and such model should be built based on learning the fertile eggs. We therefore conclude that whereas KNN classifier proved appropriate for translating the discriminative features from chicken egg fertility data into Industrial practice, handling early chicken egg fertility detection study as a one-class design problem was also shown to have very promising potentials, using fertile egg class as the learning/modelling class. Some specific one-class modelling algorithms including but not limited to OCSVM, FOCSVM, SCMPM, and one-class KNN were thereby recommended for investigation as they might offer additional learning and optimization benefits.

CONNECTING TEXT TO CHAPTER 6

Earlier works in chapters four and five have identified a lift and stable results in modelling performance using manual data resampling to tackle imbalance data problem prior further downstream analysis. Even though there have been tremendous efforts in other fields to deal with data imbalance problem, there has not been any reported research addressing this problem in the food and Agricultural sector. Chapter six of this study therefore tested the suitability of an automatic SMOTE resampling algorithm on a large chicken egg fertility data sets, in a real industrial setting, as a means of verifying online industrial adoptability of our previous modelling outcomes. Chapter 6 has been submitted to the hyperspectral imaging laboratory of the Department of Bioresource Engineering, McGill University, for proprietary reason and awaiting approval for eventual submission to the Journal of Expert Systems with Applications as:

Adegbenjo, A. O., Liu, L., and Ngadi, M. (2019). Improved Chicken Egg Fertility Classification Using SMOTE Preprocessing Algorithm and K-Nearest Neighbours' Classifier. *Expert Systems with Applications*

CHAPTER 6

DISCRIMINATING BETWEEN FERTILE AND NON-FERTILE EGGS USING 'SMOTE' PREPROCESSING ALGORITHM AND 'KNN' CLASSIFIER

Abstract

Our previous works have identified the versatility of resampling pre-processing approach in addressing the class imbalance problem with chicken egg fertility data. Even though there have been tremendous efforts in other fields to deal with data imbalance problem, no research to the best of our knowledge has been reported for Agricultural and food analysis. There is therefore a need for suitable resampling and classification algorithms that would be appropriate for an online industrial application. Oversampling and under sampling approaches have been widely reported in literatures. "Synthetic Minority Oversampling Technique" (SMOTE) is a popular algorithm, known in various field of endeavours for successfully producing additional rare class samples synthetically, as against blind duplication of samples (random oversampling by replacement). Upon the application of the SMOTE preprocessing algorithm, this paper showed the results of k-nearest neighbours classifier performance on two sets of imbalanced chicken egg fertility data, following a binary class classification problem scenario. A usable total of 9,207 white chicken egg hyperspectral imaging data, collected over a NIR wavelength range of 900-1700 nm, and on day zero (just prior to incubation), were used in the study. The data consisted of a total of 8,807 fertile eggs and 400 non-fertile eggs bringing the fertility imbalance ratio to around 1:23. The appropriateness of the SMOTE algorithm in dealing with this imbalanced problem was studied using six evaluation criteria namely: sensitivity, specificity, precision, area under ROC curve (AUC), F-measure, and overall accuracy. The analysis results showed that the SMOTE algorithm in conjunction with k-nearest neighbours' classifier is adequate in solving the imbalance problem

with chicken egg fertility classification data, and thereby presented as appropriate for building an online chicken egg fertility classification system.

6.1 Introduction

The imbalance data problem is a well-known phenomenon in various field of endeavours such as the medicals, banking, computer sciences, and text mining (Cao et al., 2015; Jindal & Liu, 2007; Phua et al., 2004; Rahman & Davis, 2013; Youn & McLeod, 2007). The problem occurs when there are higher number of instances in one category of data than in the other. During data training with various learning algorithms, the imbalance problem has been reported responsible for high predictive accuracy of the prevalent class at the expense of the minority class (Liao, 2008; Sun et al., 2009). This occurrence has made learning from an imbalanced data a critical subsector in machine learning studies (Liao, 2008). Despite there have been tremendous efforts in other fields to deal with the problem of data imbalance, no research to the best of our knowledge has been reported for Agricultural and food quality related analysis.

There have indeed been some advances towards solving the chicken egg fertility classification problem using non-destructive techniques including machine vision, spectroscopy and more recently, hyperspectral imaging (Bamelis et al., 2002; Bamelis et al., 2004; Das & Evans, 1992b; Liu & Ngadi, 2013). However, the task of dealing with the class imbalance problem associated with chicken egg fertility data and other agricultural/food analysis cases remains unaddressed. Learning challenges of imbalanced data with various standard classifiers including but not limited to neural networks, decision trees, SVM, associative classifiers, Bayesian classifiers, and KNN were well attested to in the literatures (Batista et al., 2004; Chawla et al., 2004; Japkowicz, 2000; Mani & Zhang, 2003; Sun et al., 2009), and data set pre-processing method, using resampling techniques has been found notable among others for tackling associated

learning challenges of imbalanced data (Batista et al., 2004; Batuwita & Palade, 2010; Fernández et al., 2010; Fernández et al., 2011).

Data resampling and/or pre-treatment is prerequisite for appropriate handling of imbalanced data. Resampling techniques can be grouped into three categories namely: rare class oversampling, majority class under-sampling, and hybrid (combination of oversampling and under-sampling) methods. Traditional implementation of these approaches was in terms of random under-sampling and random oversampling. Both random under-sampling and random oversampling techniques are non-heuristic in their mode of implementation. While the former technique randomly removes instances from the prevalent class (without replacement), until the number of instances in the majority class balances up with that in the minority class, the latter technique on the other hand randomly selects instances from the minority class for replication, until the number of instances in the non-prevalent class matches up with that in the prevalent class (Chawla et al., 2002; Chawla et al., 2004; Japkowicz & Stephen, 2002; Liao, 2008). This procedure is otherwise known as random oversampling with replacement. Due to the limitations usually accustomed with the aforementioned resampling approaches, ranging from loss of cogent information (that could be useful in the learning process) with random under-sampling and potential overfitting problem with random oversampling (López et al., 2013), the "Synthetic Minority Oversampling Technique" (SMOTE) algorithm, which was eventually adopted in this study was proposed by (Chawla et al., 2002), and it has become one of the most widely known resampling methods in imbalance learning.

The SMOTE algorithm creates numerous rare class synthetic examples via interpolation for oversampling instead of producing exact duplicate copies of existing examples as with random oversampling. (Liu et al., 2006) studied four sampling methods namely: oversampling with

replication, random under-sampling, ensemble under-sampling, and SMOTE to address the imbalance data distribution. It was reported that SMOTE and ensemble under-sampling methods outperformed the random under-sampling method. Similar results were reported for LDA classifier in a comparison study involving imbalanced and balanced data using under-sampling, oversampling, Tomek links, and SMOTE (Xie & Qiu, 2007). In the same vein, (López et al., 2013) observed an optimal ranking with KNN during another comparison study to evaluate performance of KNN, SVM, and C4.5 classifiers on a SMOTE balanced data. Furthermore, (Rahman & Davis, 2013) reported an improved accuracy when an imbalanced cardiovascular data was balanced using SMOTE over-sampling technique. Research has also indicated that the combination of SMOTE and under-sampling (SMOTE_RU hybrid method) does better than ordinary random under-sampling (Chawla et al., 2002), in some cases.

It was from considering the versatility of resampling approaches in combating the imbalance problem in various other field of endeavours that this study was set up to examine the applicability of the widely known "Synthetic Minority Oversampling Technique" (SMOTE), to an Agricultural/food analysis scenario. Also, its hybrid equivalent (SMOTE_RU) implemented in our study as $SMOTE_4RU$ was equally tested for handling class imbalance. Due to the bottom positioning of SMOTE generated samples, the effect of randomization procedure on improving classification accuracy with SMOTE balanced data was also simultaneously examined in this study. Chicken egg fertility data was used in conjunction with k-nearest neighbours' classifier (earlier identified as potentially appropriate for translating the research outcome of early chicken egg fertility discrimination into industrial practice). Performance evaluation was accomplished using six relevant metrics for an imbalanced data situation.

6.2 Materials and Methods

6.2.1 Samples

A total of up to 10,260 white shell eggs were acquired from a commercial hatchery. The eggs were supplied in 15 batches within a period of 16 days on an average of about 680 eggs per batch. However, only a total of 9207 eggs were used in the data analysis due to various losses mainly including egg breakage, unknown labelling information, etc. Each egg was imaged twice by manually rotating the egg 90 degrees inside its holder before a second scan. Thus, there were two different, but similar data sets of 9207 hyperspectral scans. Labelling information on fertility of the eggs was obtained upon egg break-out after about 10 days of incubation.

6.2.2 Image acquisition and processing

All the eggs were imaged prior to incubation in an industrial incubator. The eggs were handled following industrial standard procedure. The near-infrared (NIR) hyperspectral imaging system used in this project comprised of an InGaAs camera, a test rig conveyor driven by a stepping motor (C4T17FC10B, Regal Beloit America Inc., USA), a line-scan spectrograph *(*Hyperspec, Headwall Photonics Inc. USA) with a NIR spectral wavelength range from 900 to 1700 nm and a spectral resolution of 4.68 nm. A 250 W QTH lamp (Oriel Instruments, Newport Corporation, California) provided a transmission mode illumination to the eggs through an optical fibre liquid light guide. Other components of the imaging systems include frame enclosure and supporting structure, a six holed eggs holder, a data acquisition and pre-processing software *(*Hyperspec, Headwall Photonics Inc. USA) and a PC. Detailed image acquisition procedure followed is as reported by (Liu & Ngadi, 2013). The NIR wavelength region used spanned over a total of 171 wavelength bands, thereby resulting in a mean spectra data matrix of 9207 X 172

(comprising 9207 instances, 171 “independent” wavelength variables, and 1 dependent categorical variable).

6.2.3 The SMOTE algorithm and implementation

The major aim of the SMOTE algorithm as against the conventional random oversampling with replacement, is to produce new rare class instances called synthetic examples by using a distance function computed from a user defined numbers of nearest neighbours. With the SMOTE algorithm implementation, “synthetic examples” of the rare class (non-fertile eggs) were generated rather than to simply oversample by using duplicated examples. The SMOTE algorithm used in this study was implemented within the WEKA 3.9.0 platform (Frank et al., 2004; Hall et al., 2009; Witten et al., 2016), in four iterations ($SMOTE_4$) to achieve a rare class proportion comparable to the prevalent class proportion. Further information about the SMOTE algorithm and its implementation has been reported elsewhere (Chawla et al., 2002; Chawla et al., 2004). The K-nearest neighbours’ algorithm (Aha et al., 1991; Aha & Kibler, 1989) was later adopted to classify the chicken egg samples.

6.2.4 Hybrid resampling ($SMOTE_4$RU) implementation

After $SMOTE_4$ implementation, a hybrid resampling method ($SMOTE_4$RU-combining SMOTE and random under-sampling approaches) was further implemented using a ‘SpreadSubsample” under-sampling filter in WEKA to generate a random subsample of the data set. With distribution spread set at 1, a uniform maximum distribution was obtained, thereby reducing the majority class to the exact size of the minority class (Chawla et al., 2002; Frank et al., 2004; Hall et al., 2009). Other distributin ratios can be used via fine tuning of the “distributionSpread” filter.

6.2.5 Randomization procedure

Upon $SMOTE_4$ implementation, the synthetic samples generated were positioned together at the bottom, and therefore data re-ordering might be necessary to assist with appropriate class representation in the calibration, validation and testing partitions. The randomize-S 42 algorithm (Bouckaert et al., 2010) was implemented in this study, to accomplish shuffling the order of instances being passed. The algorithm consisted of a random number generator that is usually reset based on a predetermined seed value whenever a new set of samples is being passed.

6.2.6 K-nearest neighbours' classifier

K-nearest neighbours' (KNN) is among the simplest available machine learning algorithms (dated back to early 50's and late 60's), with intuitive and simple characteristic ideology (Cover & Hart, 1967; Sun et al., 2009). Upon the presentation of an unknown sample, the algorithm calculates the similarity (distance) between the new sample and all the calibration samples to determine its k-nearest neighbours, thereby deciding the class of the new sample by considering the most abundant class within the k-nearest examples (Sun et al., 2009). KNN discriminates between samples based on closest training instances in the feature learning space. An instance is classified by a majority vote of its nearest neighbours, with the sample being classified to the most common class amongst its k-nearest neighbours. For a binary classification scenario, it is a common practice choosing an odd k-value to avoid vote tie. The neighbours are usually selected from set of known correctly classified (training) samples (Shetty et al., 2010).

6.2.7 Evaluation criteria

Accuracy and performance in this study have been evaluated using six evaluation criteria namely: sensitivity (SEN), specificity (SPE), precision (PPV), area under ROC curve (AUC), F-measure (F_F and F_NF), and overall accuracy (OVA). While sensitivity (true positive rate or

recall) shows the percentage of actual fertile eggs that are predicted as fertile eggs, specificity (true negative rate) shows the percentage of actual non-fertile eggs that are predicted as non-fertile eggs. Precision (positive predictive value, PPV) on the other hand shows the percentage of predicted fertile eggs that are actual fertile eggs and vice versa with respect to the non-fertile eggs in terms of the negative predictive value, NPV. The F-measure (F1-score) has been described as the harmonic mean of precision and recall. Hence, a classifier will produce a high F1-score if both recall and precision are high. This singular metric for precision and recall speeds up decision making when considering performance of a classifier among various data types or when comparing performance of many classifiers on a data type (Andrew, 2018; Sasaki, 2007). F1-score can be manually computed using equation 1, where P is precision and R is recall. Detailed performance evaluation information for binary classification problem can be obtained from (François, 2006; He & Garcia, 2009; Prachuabsupakij, 2015; Rahman & Davis, 2013; Sun et al., 2009).

$$F1\_score = \frac{2*P*R}{P+R} \quad \text{.....6.1}$$

6.2.8 Model Validation, Testing and Verification Method

The 10-fold cross-validation approach was adopted in which the data was partitioned into 10 different segments, with each segment consisting of up to 1520 instances of the SMOTE data. A segment (of up to 1520 instances) was held out per time for validation, while the remaining 9 segments (comprising up to 1520 X 9 = 13680 instances) were used for calibration. The procedure was repeated until all the segments have successfully passed through the calibration and validation stages. Further verification was also implemented using an independent "hold-out" testing set upon data partitioning into ratio 70:30 as described in chapter 3. Model's robustness was finally tested by using independent unseen data set, containing no synthetic instances.

6.3 Results and discussion

The fertility labels from our study showed a total of 8807 fertile (F) eggs and 400 non-fertile (NF) eggs respectively, resulting in an imbalanced data distribution scenario in the ratio of up to 1NF:23F. This degree of imbalance according to (Sun et al., 2009) is more than enough to negatively affect classification accuracy. Tables 6.1 shows the quantity specifications for the overall egg samples, 70% and 30% independent sample partitions in the different data sets considered for both original and resampled egg data based on SMOTE implementation. While

Table 6.1 Quantity specifications of considered white egg samples

Data type	Overall set			70% set			30% set		
	Qty	F	NF	Qty	F	NF	Qty	F	NF
Scan1	9207	8807	400	6445	6165	280	2762	2642	120
Scan2	9207	8807	400	6445	6165	280	2762	2642	120
Scan1+Scan2	18414	17614	800	12890	12330	560	5524	5284	240
Scan1+SMOTE$_4$	15207	8807	6400	10645	6165	4480	4562	2642	1920
Scan2+SMOTE$_4$	15207	8807	6400	10645	6165	4480	4562	2642	1920
Scan1+Scan2+SMOTE$_4$	30414	17614	12800	21290	12330	8960	9124	5284	3840
Scan1+SMOTE$_4$+RU	12800	6400	6400	8960	4480	4480	3840	1920	1920
Scan2+SMOTE$_4$+RU	12800	6400	6400	8960	4480	4480	3840	1920	1920
Scan1+Scan2+SMOTE$_4$+RU	25600	12800	12800	17920	8960	8960	7680	3840	3840
Scan1+SMOTE$_4$+RAND	15207	8807	6400	10646	6211	4435	4561	2596	1965
Scan2+SMOTE$_4$+RAND	15207	8807	6400	10646	6211	4435	4561	2596	1965
Scan1+Scan2+SMOTE$_4$+RAND	30414	17614	12800	21290	12368	8922	9124	5246	3878
Scan1+SMOTE$_4$+RAND+RU	12800	6400	6400	8870	4435	4435	3930	1965	1965
Scan2+SMOTE$_4$+RAND+RU	12800	6400	6400	8870	4435	4435	3930	1965	1965
Scan1+Scan2+SMOTE$_4$+RAND+RU	25600	12800	12800	17844	8922	8922	7756	3878	3878

Figure 6.1 shows typical spectral profiles for UNSMOTE, SMOTE, and randomised SMOTE data types, where red and blue profiles represent non-fertile and fertile eggs respectively. Figure 6.2 showed typical plots for PCA sample grouping for the same sets of data (see Appendix E for the distribution profile for UNSMOTE, SMOTE, and randomised SMOTE data types. It was noticed

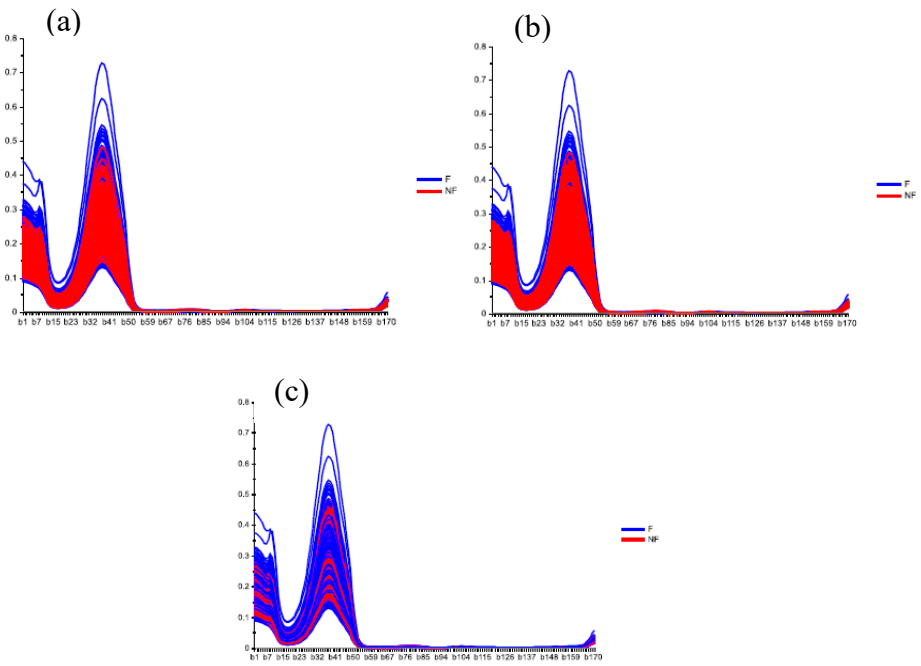


Figure 6.1: Typical spectral transmission profile (a), unsmote (b), smote (c) smote + randomization

from Figures 6.1 and 6.2 that the SMOTE profiles are much denser than the UNSMOTE profiles (Figures 6.1a, b, 6.2a, and b) due to sample population increase via SMOTE implementation. Whereas, the data point clusters in the randomised SMOTE samples were fully interwoven and embedded into one another, they were more distinguishable in the unrandomized SMOTE samples (Figures 6.1b, c, 6.2b, and c). It was observed from Figure 6.2c that usable information in the considered data are in the first principal component (1st PC) as the onset of noise is seen from PC2 (Figure 6.2d).

Tables 6.2 and 6.3 showed 10-fold cross validated results for overall and 70% data sets partitions respectively. Table 6.4 on the other hand showed the results for testing models built from 70% data on 30% unknown data. Tables 6.2, 6.3, and 6.4 showed the imbalanced data sets having the overall accuracies (OVA) ranging from 94.80% to 95.40%. However, the specificity (SPE) values were poor ranging from 0% to 3.4%. Lower AUC values ranging from 48.10% to 62.50% also showed that the reported OVA cannot be trusted being only skewed in favour of the prevalent

class, as can be seen from the sensitivity (SEN) values ranging from 99.00% to 99.70%.

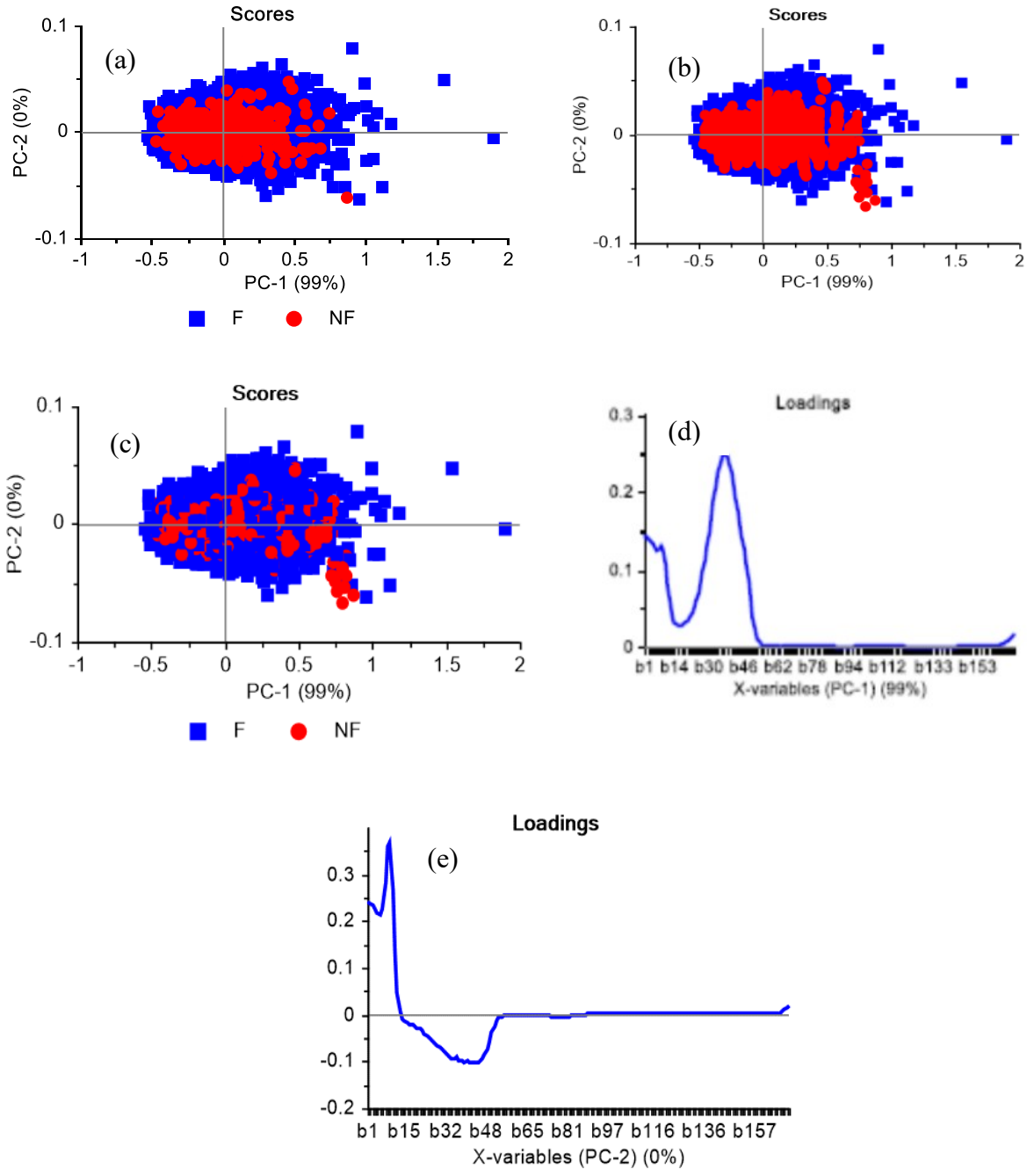


Figure 6.2: Typical PCA sample grouping plots (a), unsmote (b), smote (c) smote + randomization (d), 1st PC loading (e), 2nd PC loading

Table 6.2 3NN classifier performance (%), overall 10-fold cross validation results

Data type	SEN	SPE	PPV	NPV	AUC	F_F	F_NF	OVA
Scan1	99.50	0.30	95.60	2.30	50.90	97.50	0.50	95.20
Scan2	99.40	0.50	95.70	3.50	48.10	97.50	0.90	95.10
Scan1+Scan2	99.10	2.90	95.70	12.80	62.50	97.40	4.70	94.90
Scan1+SMOTE$_4$	93.10	97.10	97.80	91.10	97.90	95.40	94.00	94.80
Scan2+SMOTE$_4$	91.60	97.10	97.70	89.40	97.40	94.60	93.10	93.90
Scan1+Scan2+SMOTE$_4$	93.00	98.10	98.50	91.10	98.10	95.70	94.50	95.20
Scan1+SMOTE$_4$+RU	90.30	97.50	97.30	90.90	97.50	93.70	94.10	93.90
Scan2+SMOTE$_4$+RU	88.90	97.30	97.10	89.70	96.80	92.80	93.40	93.10
Scan1+Scan2+SMOTE$_4$+RU	90.10	98.30	98.20	90.90	97.50	94.00	94.40	94.20
Scan1+SMOTE$_4$+RAND	93.20	97.10	97.80	91.20	97.90	95.40	94.10	94.90
Scan2+SMOTE$_4$+RAND	91.60	97.10	97.80	89.40	97.50	94.60	93.10	94.00
Scan1+Scan2+SMOTE$_4$+RAND	93.10	98.10	98.60	91.20	98.10	95.70	94.50	95.20
Scan1+SMOTE$_4$+RAND+RU	90.10	97.50	97.30	90.80	97.40	93.60	94.00	93.80
Scan2+SMOTE$_4$+RAND+RU	88.80	97.30	97.10	89.70	96.80	92.80	93.40	93.10
Scan1+Scan2+SMOTE$_4$+RAND+RU	90.30	98.30	98.20	91.00	97.50	94.10	94.50	94.30

Table 6.3 3NN classifier percent performance: 70% 10-fold cross validated results

Data type	SEN	SPE	PPV	NPV	AUC	F_F	F_NF	OVA
Scan1	99.50	0.40	95.60	3.10	52.30	97.50	0.60	95.20
Scan2	99.20	1.40	95.70	7.50	48.50	97.40	2.40	95.00
Scan1+Scan2	99.00	3.40	95.80	13.20	62.30	97.30	5.40	94.80
Scan1+SMOTE$_4$	93.50	96.70	97.50	91.59	97.80	95.40	94.00	94.80
Scan2+SMOTE$_4$	92.60	96.90	97.60	90.50	97.70	95.10	93.60	94.40
Scan1+Scan2+SMOTE$_4$	93.00	97.90	98.40	91.00	98.20	95.60	94.30	95.10
Scan1+SMOTE$_4$+RU	91.00	97.10	97.00	91.50	97.20	93.90	94.20	94.10
Scan2+SMOTE$_4$+RU	90.20	97.10	96.90	90.90	97.30	93.50	93.90	93.70
Scan1+Scan2+SMOTE$_4$+RU	90.60	98.20	98.10	91.30	97.60	94.20	94.60	94.40
Scan1+SMOTE$_4$+RAND	91.60	95.90	96.90	89.10	97.30	94.20	92.40	93.40
Scan2+SMOTE$_4$+RAND	89.70	95.30	96.40	86.90	96.50	92.90	90.90	92.00
Scan1+Scan2+SMOTE$_4$+RAND	91.10	97.30	97.90	88.80	97.50	94.40	92.80	93.70
Scan1+SMOTE$_4$+RAND+RU	87.90	96.50	96.20	88.90	96.30	91.90	92.50	92.20
Scan2+SMOTE$_4$+RAND+RU	86.50	96.00	95.50	87.70	95.90	90.80	91.60	91.20
Scan1+Scan2+SMOTE$_4$+RAND+RU	87.70	97.60	97.30	88.80	96.60	92.30	93.00	92.70

It was observed that the $SMOTE_4$ implementation resulted in an improved recognition of the rare-class non-fertile eggs as can be seen from the obtained specificity values now ranging from 95.30% to 98.30% for all cross validated results (Tables 6.2 and 6.3). Also, AUC results improved for the same sets of $SMOTE_4$ data having values ranging from 95.90% to 98.20%. Even though cross validated results of models built from both non-randomised and randomised data sets are very similar (Tables 6.2 and 6.3), Table 6.4 showed that models from SMOTE and randomised data sets are more robust than those from SMOTE and non-randomised data sets. This was evidenced from Table 6.4 having specificity and AUC values for SMOTE and non-randomised data sets ranging from (4.60%, 47.20%) to (7.10%, 51.30%) respectively, and same set of values for SMOTE and randomised data sets ranging from (95.80%, 96.40%) to (98.00%, 97.70%), respectively. The remaining discussion of this work were based on Table 6.4 which contains the results from testing 70% cross validated models on 30% independent unknown data.

Table 6.4 3NN classifier percent performance: 30% independent testing results

Data type	SEN	SPE	PPV	NPV	AUC	F_F	F_NF	OVA
Scan1	99.70	0.00	95.60	0.00	48.90	97.60	0.00	95.30
Scan2	99.70	0.80	95.70	10.00	47.20	97.60	1.50	95.40
Scan1+Scan2	99.50	0.80	95.70	7.40	48.20	97.6 0	1.50	95.20
Scan1+$SMOTE_4$	95.60	5.30	58.10	46.60	51.30	72.30	9.50	57.60
Scan2+$SMOTE_4$	94.70	4.80	57.80	39.70	48.90	71.80	8.60	56.90
Scan1+Scan2+$SMOTE_4$	94.50	4.60	57.70	37.60	49.00	71.60	8.10	56.70
Scan1+$SMOTE_4$+RU	93.90	7.10	50.30	53.90	51.30	65.50	12.60	50.50
Scan2+$SMOTE_4$+RU	91.30	6.20	49.30	41.60	47.20	64.00	10.80	48.80
Scan1+Scan2+$SMOTE_4$+RU	92.70	6.20	49.70	45.90	48.40	64.70	10.90	49.50
Scan1+$SMOTE_4$+RAND	92.30	95.80	96.70	90.40	97.20	94.40	93.00	93.80
Scan2+$SMOTE_4$+RAND	91.20	96.20	97.00	89.20	97.00	94.00	92.60	93.40
Scan1+Scan2+$SMOTE_4$+RAND	92.10	97.60	98.10	90.10	97.70	95.00	93.70	94.40
Scan1+$SMOTE_4$+RAND+RU	88.70	96.80	96.50	89.50	96.80	92.40	93.00	92.70
Scan2+$SMOTE_4$+RAND+RU	87.60	96.60	96.30	88.60	96.40	91.70	92.40	92.10
Scan1+Scan2+$SMOTE_4$+RAND+RU	89.20	98.00	97.80	90.10	97.10	93.30	93.90	93.60

From the ROC curve plotted in Figure 6.3, it was observed that all six data types (10-15) above the random line, are all SMOTE and randomised data. These data types therefore possess the best structure exposable to the KNN learning algorithm. In a ROC curve, the best model is

usually positioned at the top left-hand corner of the ROC curve. This made the models built from data types 10, 11, and 12 (Scan1+SMOTE$_4$+RAND, Scan2+SMOTE$_4$+RAND, and Scan1+Scan2+SMOTE$_4$+RAND) much preferable than those built from data types 13, 14, and 15 (which are hybrid resampling and randomised based data types). This observation also showed that the hybrid resampling technique (SMOTE$_4$RU) does not necessarily improve the KNN classifier performance as the positions of all models with RU implementation were below those without RU implementation in the ROC space. This observation is similar to that of (Rahman & Davis, 2013), in which the under-sampling approach adopted did not improve accuracy and sensitivity values of the medical data sets considered.

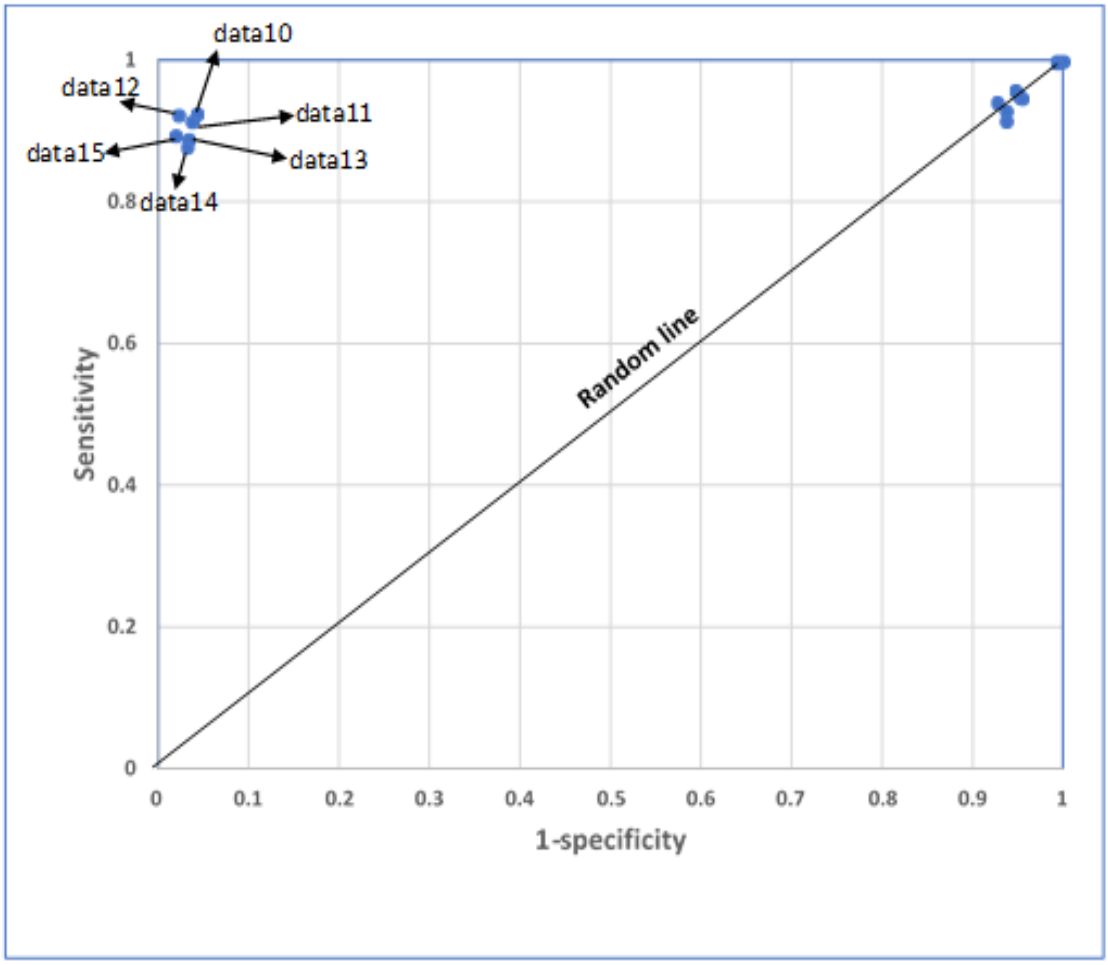


Figure 6.3: ROC of 3NN classifier

Figure 6.4 showed other evaluation metric curves which are: F-score, AUC and precision/recall curves. Even though the F-score (F_F) for models 1, 2, and 3 were high for fertile eggs (all 97.60%) as shown in Fig.6.4a, these models cannot be taken as great because F-score values for the same models with respect to the non-fertile eggs were low (0%, 1.50%, and 1.50%). The reason for this outcome being that precision/recall for fertile eggs were high while precision/recall for non-fertile eggs were low due to ratio imbalance. A better informative decision might therefore be taken considering the average F-score (Andrew, 2018) and AUC plots as shown in Figure 6.4b. Models 10-15 from data types 10-15 earlier observed from the ROC analysis also

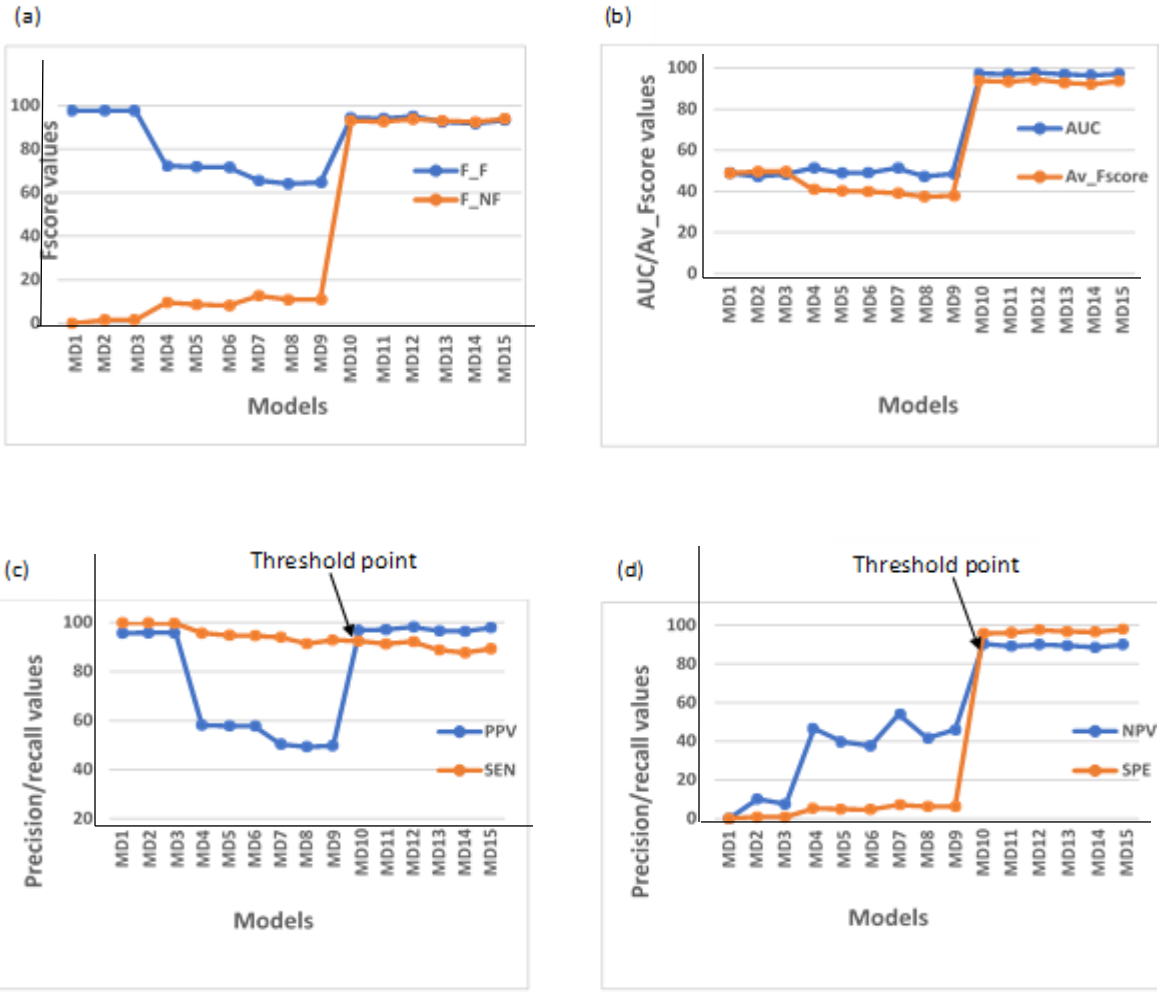


Figure 6.4: Evaluation metric curves (a), F-score (b), AUC/Average F-score (c), Precision/recall curve for fertile eggs (d), Precision/recall curve for non-fertile eggs

top the list for both average F-score and AUC values (Figure 6.4b). The best three models however were models 12, 10, and 15 having respectively average F-score and AUC values of (94.35%, 97.70%); (93.70%, 97.20%); and (93.60%, 97.10%). It was noticed that model built from data 11 alone was not as robust as that built from data 12 (a combination of data 10 and 11). While random under-sampling does not necessarily improve classifier performance as can be seen from models 13 and 14, it seems to offer some improvement with more data addition as can be seen with model 15 when data 13 and 14 were combined.

A notable challenge with the combination of precision and recall values into a singular metric F-score is the precision/recall trade-off. Generally, an increasing recall do result in a decreasing precision and vice versa. This is therefore a problem in a situation where one is interested in both high precision and high recall. Hence, the need to determine an appropriate threshold point at which precision and recall values could be acceptable for the choice model. This problem was addressed in the precision/recall curves for fertile and non-fertile eggs as shown in Figures 6.4c and 6.4d. Therefore, for best recognition of fertile and non-fertile eggs, model 10 built from smote and randomized data type 10 was chosen as the optimal model, considering the threshold point from the precision/recall curves.

Finally, to have the real idea of our model performance on future unknown data, the models 10, 12, and 15 built from 70% data were tested on their corresponding 30% independent data but excluding all synthetic samples as futuristic unseen data will not contain synthetic samples. This step made our verification data to be consistent with the suggestion of (Kuhn & Johnson, 2016), that testing test must be drawn to reflect the imbalanced state of futuristic data. Table 6.5 showed the confusion matrices for this final verification test alongside their counterpart confusion matrices with synthetic samples included, for comparison purpose. The confusion matrix has the true class

in the rows and the predicted class in the columns (Table 6.5a). Also, positive class were taken as the fertile eggs (F), negative class as the non-fertile eggs (NF), true positive as TP, true negative as TN, false positive as FP, and false negative as FN. The evaluation metrics for the 30% independent data, but excluding synthetic samples was finally presented in Table 6.6, for confusion matrix Table 6.5 d, f, and h.

It was observed from the confusion matrices shown that the $SMOTE_4$ models (Table 6.5 c to h), outperform the unsmote model (Table 6.5b). While the unsmote model failed totally in classifying correctly the rare class non-fertile eggs (0% recognition rate), the recognition rate for $SMOTE_4$ models (even without synthetic samples) ranges from 58.90% to 83.50% (Table 6.5 d, f, and h). This is also clearly shown in the metric evaluation Table 6.6b in terms of the specificity values.

It was further observed that testing set with synthetic samples (Table 6.5c), has 201 fertile eggs misclassified as non-fertile eggs out of a total 2596 fertile eggs and 83 non-fertile eggs misclassified as fertile eggs out of a total 1965 non-fertile eggs. With a total of 1853 identified synthetic non-fertile eggs removed from this testing set, similar results with respect to the fertile eggs were obtained in Table 6.5d, having 197 fertile eggs misclassified as non-fertile eggs (4 more fertile eggs gain), and 66 non-fertile eggs correctly classified as non-fertile out of 112 real non-fertile eggs tested. This result inferred that a total of 1816 samples from the 1882 correctly classified non-fertile eggs (Table 6.5c) were synthetic non-fertile eggs, since only 66 were eventually correctly classified after removing the synthetic samples from the testing set. This also inferred that 37 out of the 83 misclassified non-fertile eggs were synthetic as indeed only 46 real non-fertile eggs were eventually misclassified. These results were improved with more data addition as can be seen in model 12 confusion matrices (Table 6.5 e and f). It is also crucial to note

that this improvement in non-fertile egg recognition is not at any serious detriment of fertile egg recognition as can be seen from sensitivity (SEN) and precision (PPV) values in Table 6.6b, ranging from (89.10% to 92.40%) and (98.10% to 99.10%).

Table 6.5 Confusion matrix for selected models (a), Typical confusion matrix table (b), Unsmote Model 1 (c), Model 10 with synthetic (d), Model 10 without synthetic (e), Model 12 with synthetic (f), Model 12 without synthetic (g), Model 15 with synthetic (h), Model 15 without synthetic

(a)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	TP	FN
	Actually Negative (NF)	FP	TN

(b)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	2633	9
	Actually Negative (NF)	120	0

(c)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	2395	201
	Actually Negative (NF)	83	1882

(d)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	2399	197
	Actually Negative (NF)	46	66

(e)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	4829	417
	Actually Negative (NF)	92	3786

(f)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	4829	417
	Actually Negative (NF)	46	190

(g)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	3460	418
	Actually Negative (NF)	79	3799

(h)

		Prediction class	
		Predicted Positive (F)	**Predicted Negative (NF)**
True class	**Actually Positive (F)**	3456	422
	Actually Negative (NF)	39	197

Table 6.6 3NN classifier performance: 30% independent verification testing (without synthetic samples) (a), Sample specification (b), Evaluation metric results

(a)

Data type	Model	Quantity	F	NF	NF_Synth
Scan1(Unsmote)	1	2762	2642	120	-
Scan1+SMOTE4+RAND	10	2708	2596	112	1853
Scan1+Scan2+SMOTE4+RAND	12	5482	5246	236	3642
Scan1+Scan2+SMOTE4+RAND+RU	15	4114	3878	236	3642

(b)

Data type	SEN	SPE	PPV	NPV	AUC	F_F	F_NF	ACC
Scan1	99.70	0.00	95.60	0.00	48.90	97.60	0.00	95.30
Scan1+SMOTE$_4$+RAND	**92.40**	**58.90**	**98.10**	25.10	**81.80**	**95.20**	35.20	91.00
Scan1+Scan2+SMOTE$_4$+RAND	**92.10**	**80.50**	**99.10**	31.30	**91.70**	**95.40**	45.10	91.60
Scan1+Scan2+SMOTE$_4$+RAND+RU	**89.10**	83.50	**98.90**	31.80	91.50	93.70	46.10	88.80

6.4 Conclusion

This study has shown that the implementation of the SMOTE algorithm on randomised chicken egg fertility data possesses great potential of helping the minority class learning. The study also showed that increment in sample population size improved classification accuracy. Results were generally great with SMOTE implementation in comparison to without SMOTE implementation. Caution must however be taken in interpreting SMOTE results as prediction can be largely skewed towards predicting synthetic rare class samples as against real rare class samples. This possibility seems inevitable when you have too much synthetic rare class data in comparison to real rare class data. It is also crucial to make sure that the testing sample (so far, any form of resampling has been applied on calibration set), be drawn to reflect the imbalanced state of real unseen data for better futuristic model generalisation.

Future work will be geared towards improving algorithm learning of the rare class and at the same time optimising the performance on the majority class. Efforts would be made to test other forms of SMOTE iterations to reduce quantity of synthetic samples and at the same time optimising rare class data learning. Cost sensitive learning, appropriate data randomisation, and hybrid systems of oversampling and under-sampling seem to have promising potentials to help in optimising the present results. Also testing other classifiers apart from the K-nearest neighbours on SMOTE data, and fine tuning of classifier parameters may also proffer additional optimisation benefits.

CONNECTING TEXT TO CHAPTER 7

Chapters three through six have identify gaps limiting having robust modelling structure for early chicken egg fertility prediction, and proffered solutions for closing such gaps. This final chapter 7 of the thesis provided the summary of all results in the earlier chapters. In the same vein, it also presented the contribution this study has made both to the scientific community and the hatchery industries. The chapter concluded with further research recommendations towards improving stability and reproducibility of built models, and thereby improving verification outcome of chicken egg fertility predictive modelling.

CHAPTER 7

GENERAL CONCLUSION AND RECOMMENDATIONS

7.1 General conclusion

The problem of early chicken egg fertility discrimination is a very huge challenge for the hatchery industries, resulting in loss of millions of dollars annually. Conventional method of chicken egg fertility assessment termed candling is subjective, cumbersome, slow, and inefficient. Apart from the candling system being laborious and inaccurate, it is also not appropriate for building an online egg fertility classification system, in a fast pace technology advancing era of the day. Hence, a non-destructive hyperspectral imaging-based technology have been used to develop robust predictive models to assist with early chicken egg fertility identification problem.

The work commenced by investigating the possibility of using a supervised learning PLS regression algorithm for building predictive discriminative model for early chicken egg fertility assessment. Moving thresholding technique was implemented in conjunction with varying number of PLS components (PCs) ranging from 5 to 50. Results' analysis showed that the adapted PLS regression algorithm can accurately discriminate both brown and white fertile eggs from non-fertile eggs, on various incubation periods ranging from day 0 to day 4. Hence, presenting chicken egg fertility data as having discriminative information, even prior to incubation. To take advantage of using fewer number of PCs for modelling however, the PLS regression implementation mode was recommended for improvement due to imbalanced data structure existing in the considered data sets. Therefore, the subsequent study evaluated the suitability of using a ratio-based feature selection technique, towards improving model performance for early chicken egg fertility discrimination. A PLSDA learner-based feature selection algorithm, in conjunction with a non-

parametric ROC curve analysis approach was confirmed appropriate for identifying important features towards early prediction of chicken egg fertility. With optimal models presented mostly rated excellent from AUC values in the range of 90-100%, chicken egg fertility model structure was eventually successfully developed, validated, and verified on resampled data sets, using adequate number of PCs.

The study progressed by examining various other classification algorithms (including PLSDA), to ascertain which would best expose chicken egg fertility data structure to learning algorithms towards building an industrial online classification system. KNN classifier was found preferable among all other classifiers considered and was therefore used to assess the potential of a SMOTE resampling algorithm in improving chicken egg fertility classification accuracies, using a fairly-large industrial data set. The final results showed that the SMOTE algorithm was adequate in lifting models' performance. With sensitivity = 92.10%, specificity = 80.50%, precision = 99.10%, F1-score = 95.40%, AUC = 91.70%, and overall accuracy = 91.60%, our model structure was considered stable and ready for preliminary industrial testing, having been verified using independent and non-synthetic imbalanced data.

7.2 Contribution to knowledge

This work made original contributions towards strengthening existing knowledge of a great potential of hyperspectral imaging technique in an objective assessment of food products. The techniques developed from this study will find industrial applications in the future real time and online detection/classification systems for food products.

1. For the first time, an automatic synthetic minority oversampling technique (SMOTE) was successfully used to handle imbalanced data problem in an Agricultural/food processing application scenario.

2. A ratio-based metabolomic feature selection approach was for the first time introduced to identify informative variables (features) and thereby selecting appropriate discriminating features from chicken egg fertility data.
3. For the first time in chicken egg fertility studies, classification model structures for day 0 incubation were successfully developed, validated, and verified on imbalanced data. Models are ready for preliminary industrial testing. Our gold standard result is as presented thus:
    A. SMOTE best model verification performance for white eggs: PPV (Precision) = 99.10%; SEN (Sensitivity or recall) = 92.10%; SPE (Specificity) = 80.50%; F1-score = 95.40%; AUC = 91.70%, ACC (Overall accuracy) = 91.60%.
    B. SMOTE worst model verification performance for white eggs: PPV (Precision) = 98.10%; SEN (Sensitivity or recall) = 92.40%; SPE (Specificity) = 58.90%; F1-score = 95.20%; AUC = 81.80%, ACC (Overall accuracy) = 91.00%.
    C. Ratio Model best verification performance for both brown and white eggs: “hold-out” prediction accuracy = 100%
    D. Ratio Model worst verification performance for white eggs: “hold-out” prediction accuracy = 95%
4. For the first time, supervised learning PLS regression algorithm was successfully used adaptively, for chicken egg fertility classification task, prior to incubation.
5. This study made available scientific data that can be grown into a big data bank for future training and retraining of machine learning models for chicken egg fertility assessment.

7.3 General Recommendation

Other modes of SMOTE implementation, peradventure in conjunction with some available signal quality processing algorithms needs to be investigated to uplift present performance. There is a need to integrate the ratio feature selection and SMOTE approaches for possible results upliftment, towards building an industrial multispectral system. More works need to be done with larger data sets to ensure stability and reproducibility of present results. Other researchers (including from our lab) in the global machine learning and artificial intelligence community are encouraged to work more in uplifting the present results towards a long-awaited paradigm shift in the hatchery industries.

APPENDICES

Appendix Table A1. Percent classification accuracy for brown eggs based on 25 PLS components

INC. DAY	TR	FP	FN	TP	TN	TPR (%)	TNR (%)	OVA (%)
Day 0	0.5	1	8	15	311	65.22	99.68	97.31
F = 312	0.55	1	3	20	311	86.96	99.68	98.81
NF = 23	0.60	1	3	20	311	86.96	99.68	98.81
T = 335	0.65	3	3	20	309	86.96	99.04	98.21
	0.70	12	3	20	300	86.96	96.15	95.52
	0.75	14	1	22	298	95.65	95.51	95.52
	0.80	29	1	22	283	95.65	90.71	91.04
	0.81	30	0	23	282	100	90.38	91.04
Day 1	0.5	0	10	13	311	56.52	100	97.01
F = 311	0.55	0	8	15	311	65.22	100	97.6
NF = 23	0.60	1	4	19	310	82.61	99.68	98.5
T = 334	0.65	3	2	21	308	91.3	99.04	98.5
	0.70	6	1	22	305	95.65	98.07	97.9
	0.75	16	1	22	295	95.65	94.86	94.91
	0.79	32	1	22	279	95.65	89.71	90.12
	0.80	37	0	23	274	100	88.1	88.92
Day 2	0.5	0	4	19	311	82.61	100	98.8
F = 311	0.55	0	3	20	311	86.96	100	99.1
NF = 23	0.60	0	1	22	311	95.65	100	99.7
T = 334	0.65	0	1	22	311	95.65	100	99.7
	0.70	2	1	22	309	95.65	99.36	99.1
	0.75	8	1	22	303	95.65	97.43	97.31
	0.79	18	1	22	293	95.65	94.21	94.31
	0.80	21	1	22	290	95.65	93.25	93.41
Day 3	0.5	0	7	16	311	69.57	100	97.9
F = 311	0.55	0	7	16	311	69.57	100	97.9
NF = 23	0.60	0	6	17	311	73.91	100	98.2
T = 334	0.65	1	4	19	310	82.61	99.68	98.5
	0.70	5	3	20	306	86.96	98.39	97.6
	0.75	15	2	21	296	91.3	95.18	94.91
	0.79	25	0	23	286	100	91.96	92.51
	0.80	28	0	23	283	100	91	91.62
Day 4	0.5	0	6	17	311	73.91	100	98.2
F = 311	0.55	0	5	18	311	78.26	100	98.5
NF = 23	0.60	0	5	18	311	78.26	100	98.5
T = 334	0.65	1	3	20	310	86.96	99.68	98.8
	0.70	4	0	23	307	100	98.71	98.8
	0.75	13	0	23	298	100	95.82	96.11
	0.79	17	0	23	294	100	94.53	94.91
	0.80	20	0	23	291	100	93.57	94.01

Appendix Table A2. Percent classification accuracy for white eggs based on 25 PLS components

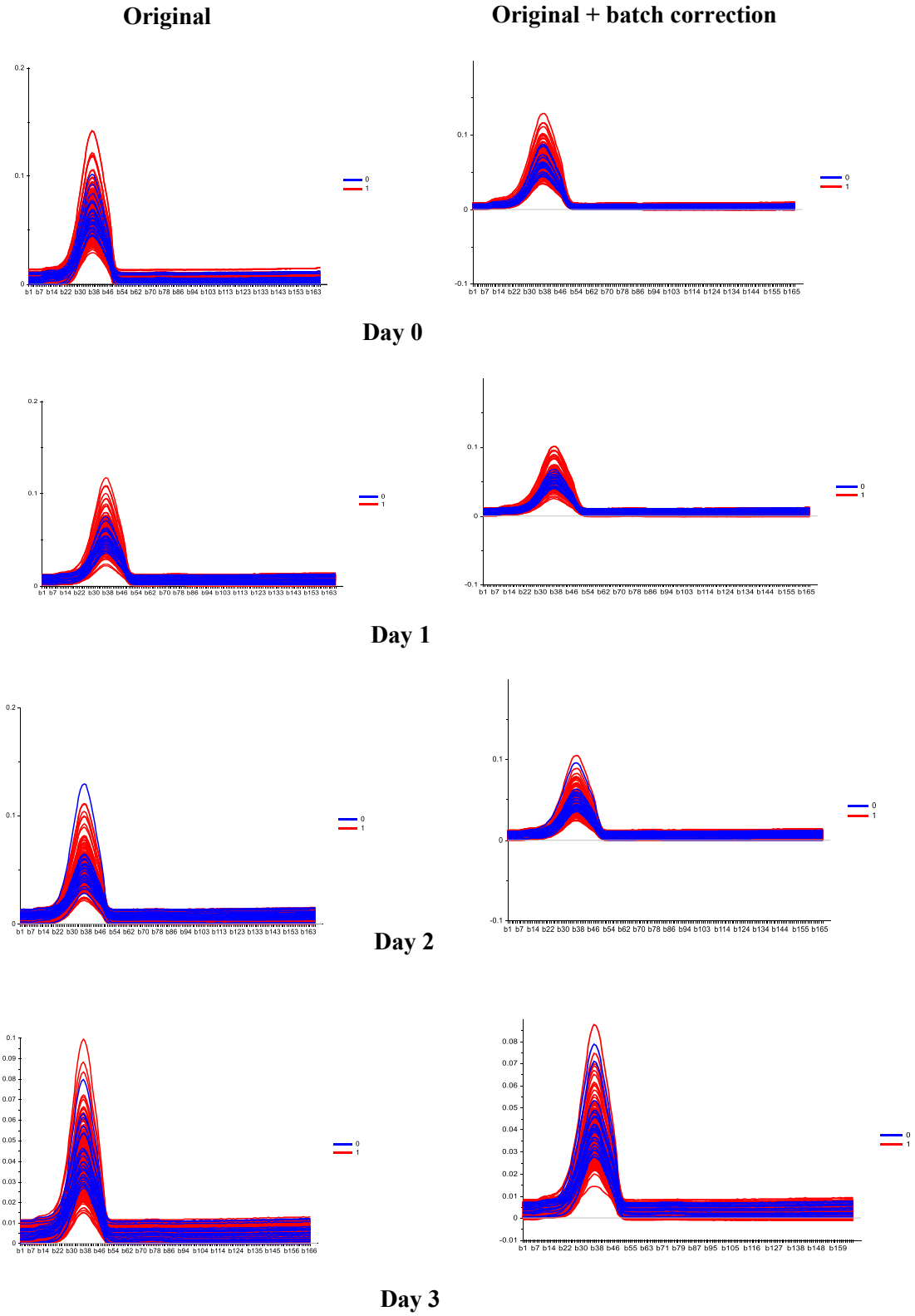
INC. DAY	TR	FP	FN	TP	TN	TPR (%)	TNR (%)	OVA (%)
Day 0	0.5	0	10	11	314	52.38	100	97.01
F = 314	0.55	0	8	13	314	61.9	100	97.61
NF = 21	0.60	0	7	14	314	66.67	100	97.91
T = 335	0.65	0	6	15	314	71.43	100	98.21
	0.70	2	3	18	312	85.71	99.36	98.51
	0.75	7	3	18	307	85.71	97.77	97.01
	0.80	21	1	20	293	95.24	93.31	93.43
	0.81	23	1	20	291	95.24	92.68	92.84
Day 1	0.5	0	5	16	314	76.19	100	98.51
F = 314	0.55	0	3	18	314	85.71	100	99.1
NF = 21	0.60	1	3	18	313	85.71	99.68	98.81
T = 335	0.65	2	2	19	312	90.48	99.36	98.81
	0.70	6	2	19	308	90.48	98.09	97.61
	0.75	10	1	20	304	95.24	96.82	96.72
	0.80	21	1	20	293	95.24	93.31	93.43
	0.81	26	1	20	288	95.24	91.72	91.94
Day 2	0.5	0	8	13	314	61.9	100	97.61
F = 314	0.55	0	5	16	314	76.19	100	98.51
NF = 21	0.60	0	3	18	314	85.71	100	99.1
T = 335	0.65	2	1	20	312	95.24	99.36	99.1
	0.70	3	1	20	311	95.24	99.04	98.81
	0.75	10	1	20	304	95.24	96.82	96.72
	0.80	26	0	21	288	100	91.72	92.24
	0.81	31	0	21	283	100	90.13	90.75
Day 3	0.5	0	4	17	312	80.95	100	98.8
F = 312	0.55	0	3	18	312	85.71	100	99.1
NF = 21	0.60	0	3	18	312	85.71	100	99.1
T = 333	0.65	1	2	19	311	90.48	99.68	99.1
	0.70	3	2	19	309	90.48	99.04	98.5
	0.75	10	1	20	302	95.24	96.79	96.7
	0.79	23	0	21	289	100	92.63	93.09
	0.80	27	0	21	285	100	91.35	91.89
Day 4	0.5	0	5	16	312	76.19	100	98.5
F = 312	0.55	0	5	16	312	76.19	100	98.5
NF = 21	0.60	1	3	18	311	85.71	99.68	98.8
T = 333	0.65	2	0	21	310	100	99.36	99.4
	0.70	3	0	21	309	100	99.04	99.1
	0.75	5	0	21	307	100	98.4	98.5
	0.79	12	0	21	300	100	96.15	96.4
	0.80	13	0	21	299	100	95.83	96.1

Appendix Table A3. Percent classification accuracy for brown eggs based on 5 PLS components

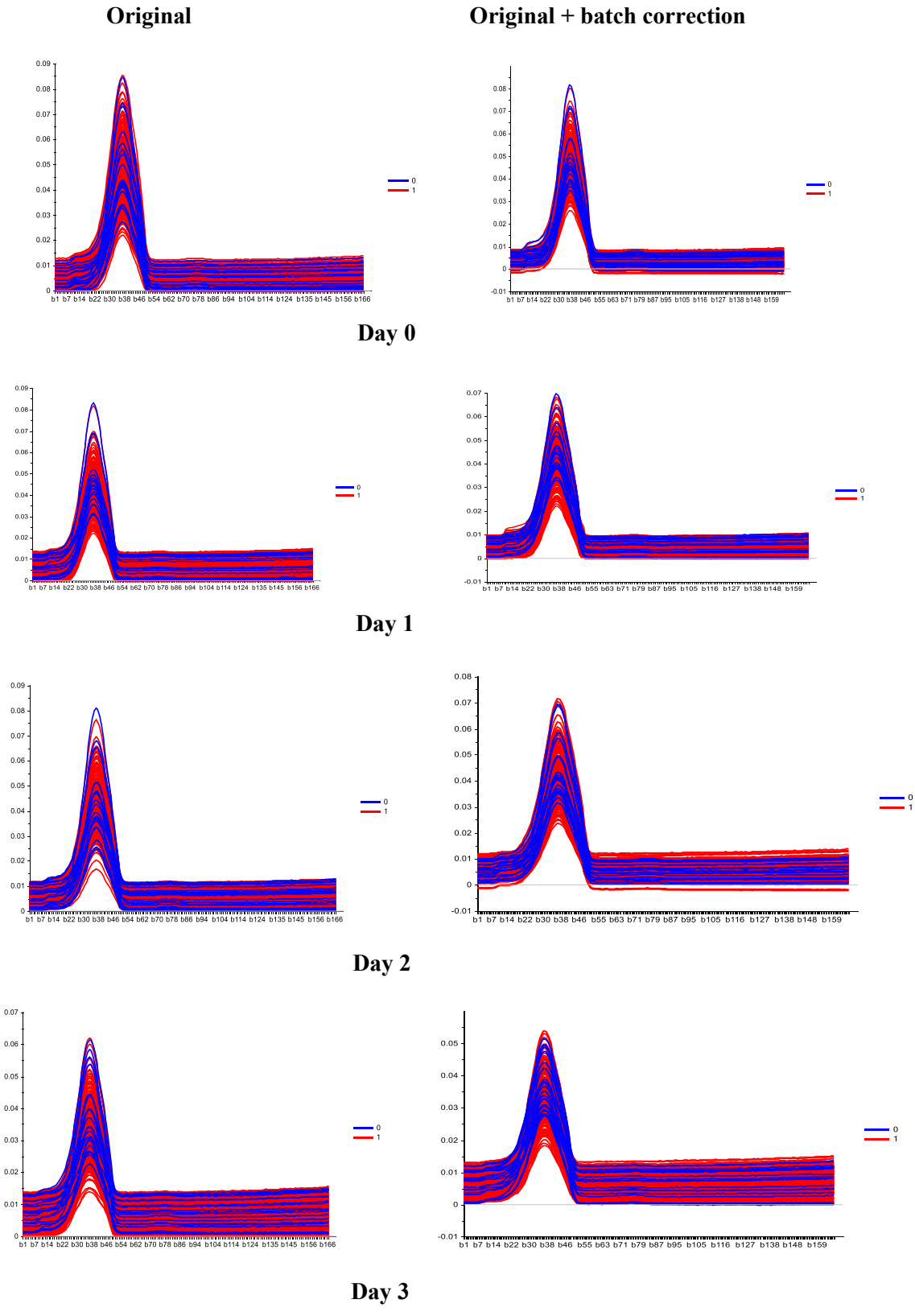
INC. DAY	TR	FP	FN	TP	TN	TPR (%)	TNR (%)	OVA (%)
Day 0	0.5	0	23	0	312	0	100	93.13
F = 312	0.55	0	23	0	312	0	100	93.13
NF = 23	0.60	0	23	0	312	0	100	93.13
T = 335	0.65	0	23	0	312	0	100	93.13
	0.70	0	23	0	312	0	100	93.13
	0.75	0	23	0	312	0	100	93.13
	0.80	0	23	0	312	0	100	93.13
	0.81	0	23	0	312	0	100	93.13
Day 1	0.5	0	23	0	311	0	100	93.11
F = 311	0.55	0	23	0	311	0	100	93.11
NF = 23	0.60	0	23	0	311	0	100	93.11
T = 334	0.65	0	23	0	311	0	100	93.11
	0.70	0	23	0	311	0	100	93.11
	0.75	0	23	0	311	0	100	93.11
	0.79	0	23	0	311	0	100	93.11
	0.80	0	23	0	311	0	100	93.11
Day 2	0.5	0	23	0	311	0	100	93.11
F = 311	0.55	0	23	0	311	0	100	93.11
NF = 23	0.60	0	23	0	311	0	100	93.11
T = 334	0.65	0	23	0	311	0	100	93.11
	0.70	0	23	0	311	0	100	93.11
	0.75	0	22	1	311	4.35	100	93.41
	0.79	1	22	1	310	4.35	99.68	93.11
	0.80	1	22	1	310	4.35	99.68	93.11
Day 3	0.5	0	23	0	311	0	100	93.11
F = 311	0.55	0	22	1	311	4.35	100	93.41
NF = 23	0.60	1	22	1	310	4.35	99.68	93.11
T = 334	0.65	1	22	1	310	4.35	99.68	93.11
	0.70	1	22	1	310	4.35	99.68	93.11
	0.75	1	20	3	310	13.04	99.68	93.71
	0.79	4	19	4	307	17.39	98.71	93.11
	0.80	4	19	4	307	17.39	98.71	93.11
Day 4	0.5	0	20	3	311	13.04	100	94.01
F = 311	0.55	1	19	4	310	17.39	99.68	94.01
NF = 23	0.60	1	18	5	310	21.74	99.68	94.31
T = 334	0.65	1	17	6	310	26.09	99.68	94.61
	0.70	2	17	6	309	26.09	99.36	94.31
	0.75	2	15	8	309	34.78	99.36	94.91
	0.79	5	15	8	306	34.78	98.39	94.01
	0.80	8	14	9	303	39.13	97.43	93.41

Appendix Table A4. Percent classification accuracy for white eggs based on 5 PLS components

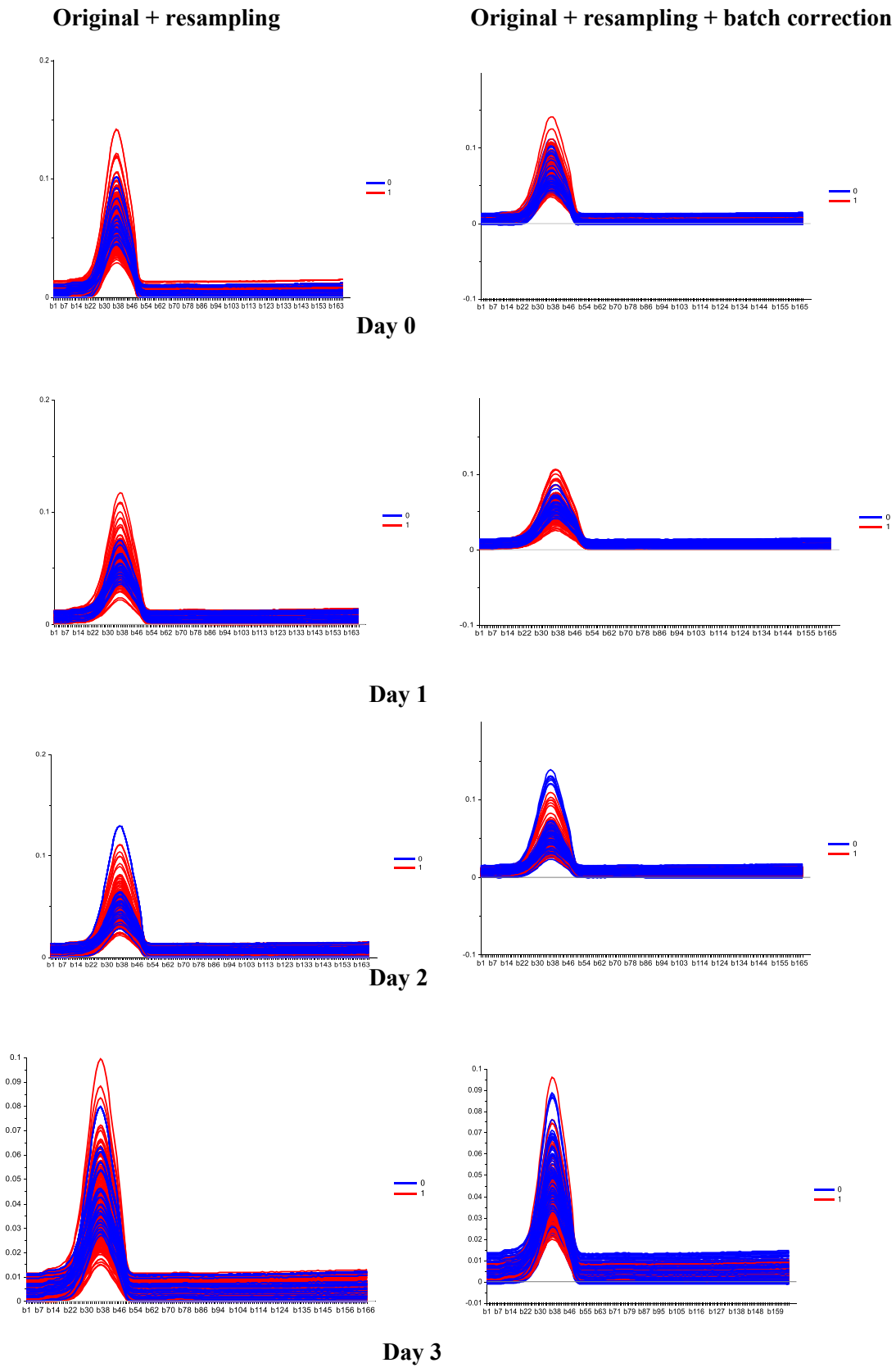
INC. DAY	TR	FP	FN	TP	TN	TPR (%)	TNR (%)	OVA (%)
Day 0	0.5	0	21	0	314	0	100	93.73
F = 314	0.55	0	21	0	314	0	100	93.73
NF = 21	0.60	0	21	0	314	0	100	93.73
T = 335	0.65	0	21	0	314	0	100	93.73
	0.70	0	21	0	314	0	100	93.73
	0.75	0	21	0	314	0	100	93.73
	0.80	3	20	1	311	4.76	99.04	93.13
	0.81	4	20	1	310	4.76	98.73	92.84
Day 1	0.5	0	21	0	314	0	100	93.73
F = 314	0.55	0	21	0	314	0	100	93.73
NF = 21	0.60	0	21	0	314	0	100	93.73
T = 335	0.65	0	21	0	314	0	100	93.73
	0.70	0	21	0	314	0	100	93.73
	0.75	0	21	0	314	0	100	93.73
	0.80	1	21	0	313	0	99.68	93.43
	0.82	3	21	0	311	0	99.04	92.84
Day 2	0.5	0	21	0	314	0	100	93.73
F = 314	0.55	0	21	0	314	0	100	93.73
NF = 21	0.60	0	21	0	314	0	100	93.73
T = 335	0.65	0	21	0	314	0	100	93.73
	0.70	0	21	0	314	0	100	93.73
	0.75	0	21	0	314	0	100	93.73
	0.80	1	20	1	313	4.76	99.68	93.73
	0.82	5	18	3	309	14.29	98.41	93.13
Day 3	0.5	0	21	0	312	0	100	93.69
F = 312	0.55	0	21	0	312	0	100	93.69
NF = 21	0.60	0	21	0	312	0	100	93.69
T = 333	0.65	0	21	0	312	0	100	93.69
	0.70	1	20	1	311	4.76	99.68	93.69
	0.75	2	19	2	310	9.52	99.36	93.69
	0.79	7	18	3	305	14.29	97.76	92.49
	0.80	8	18	3	304	14.29	97.44	92.19
Day 4	0.5	1	14	7	311	33.33	99.68	95.50
F = 312	0.55	1	14	7	311	33.33	99.68	95.50
NF = 21	0.60	1	13	8	311	38.10	99.68	95.80
T = 333	0.65	4	10	11	308	52.38	98.72	95.80
	0.70	5	10	11	307	52.38	98.40	95.50
	0.75	8	7	14	304	66.67	97.44	95.50
	0.79	10	7	14	302	66.67	96.79	94.89
	0.80	12	7	14	300	66.67	96.15	94.29



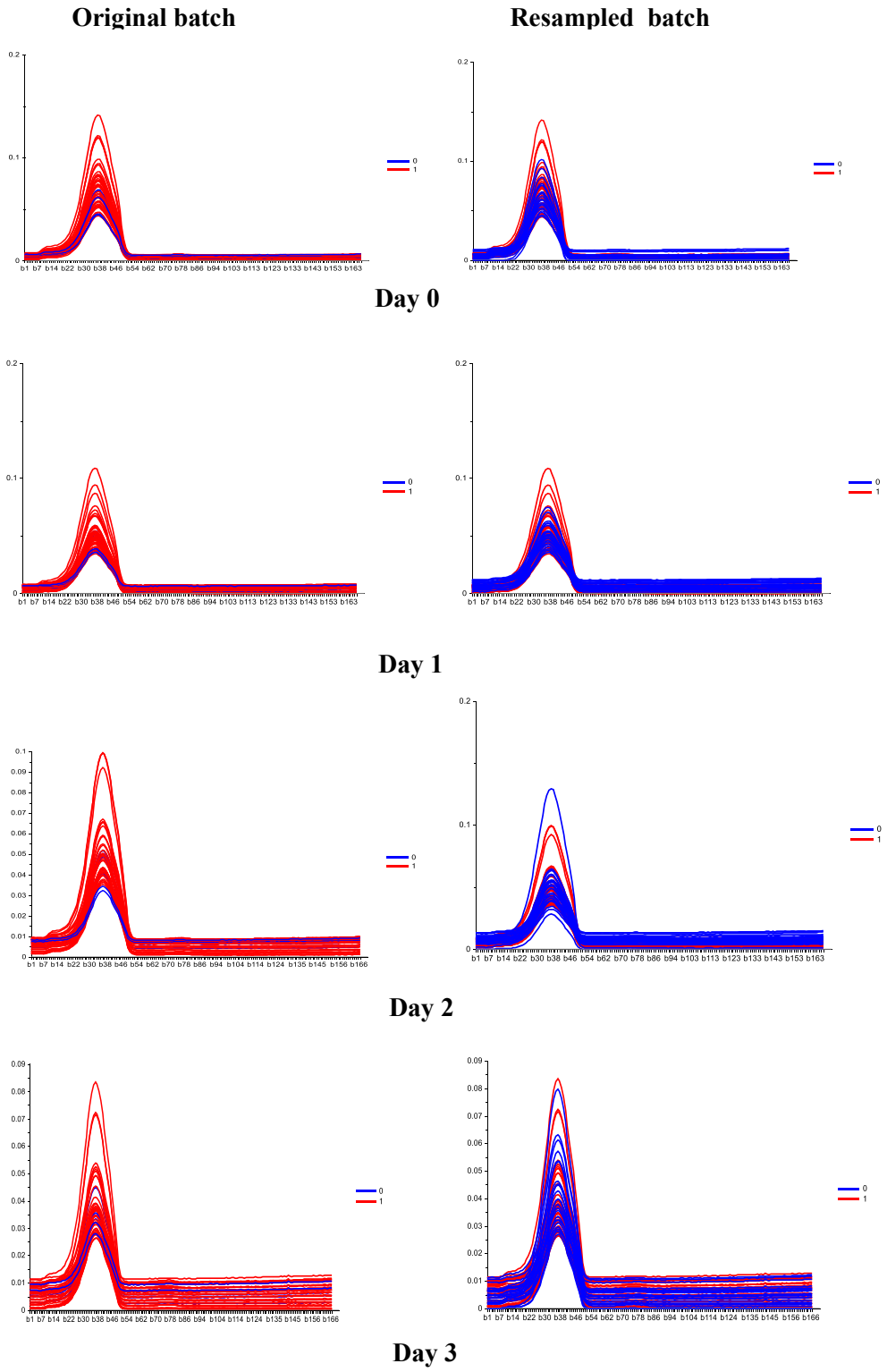
Appendix Figure B1: Typical spectral transmission profile for original and batch corrected brown egg data



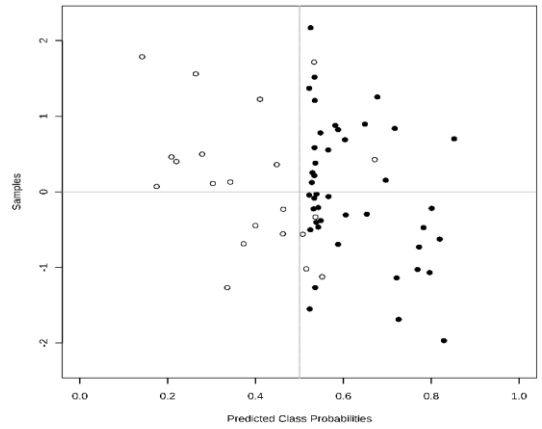
Appendix Figure B2: Typical spectral transmission profile for original and batch corrected white egg data



Appendix Figure B3: Typical spectral transmission profile for resampled and batch corrected egg data



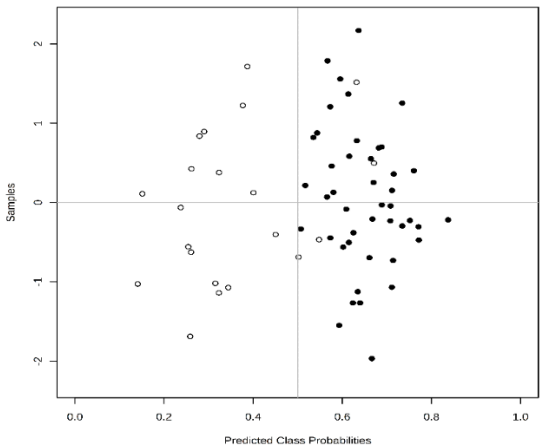
Appendix Figure B4: Typical spectral transmission profile for original and resampled batch egg data



True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	100 (43/43)	0 (0)
Actually Negative	28.57 (6/21)	71.43 (15/21)

OVA = 90.62%

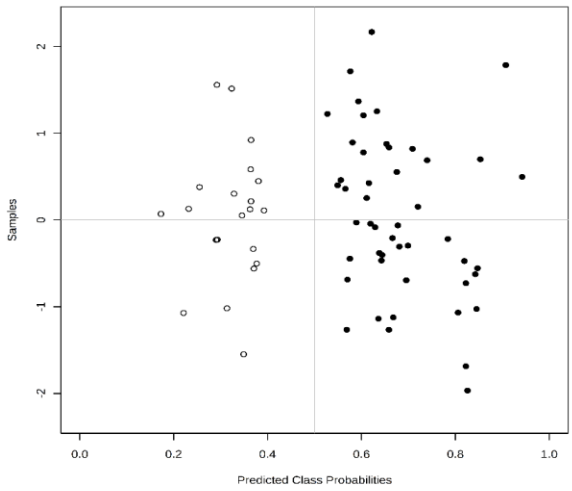
Least model



True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	100 (44/44)	0 (0)
Actually Negative	19.05 (4/21)	80.95 (17/21)

OVA = 93.84%

Median model



True class (%)	Prediction class (%)	
	Predicted Positive	**Predicted Negative**
Actually Positive	100 (48/48)	0 (0)
Actually Negative	0(0)	100 (21/21)

OVA = 100.00%

Best model

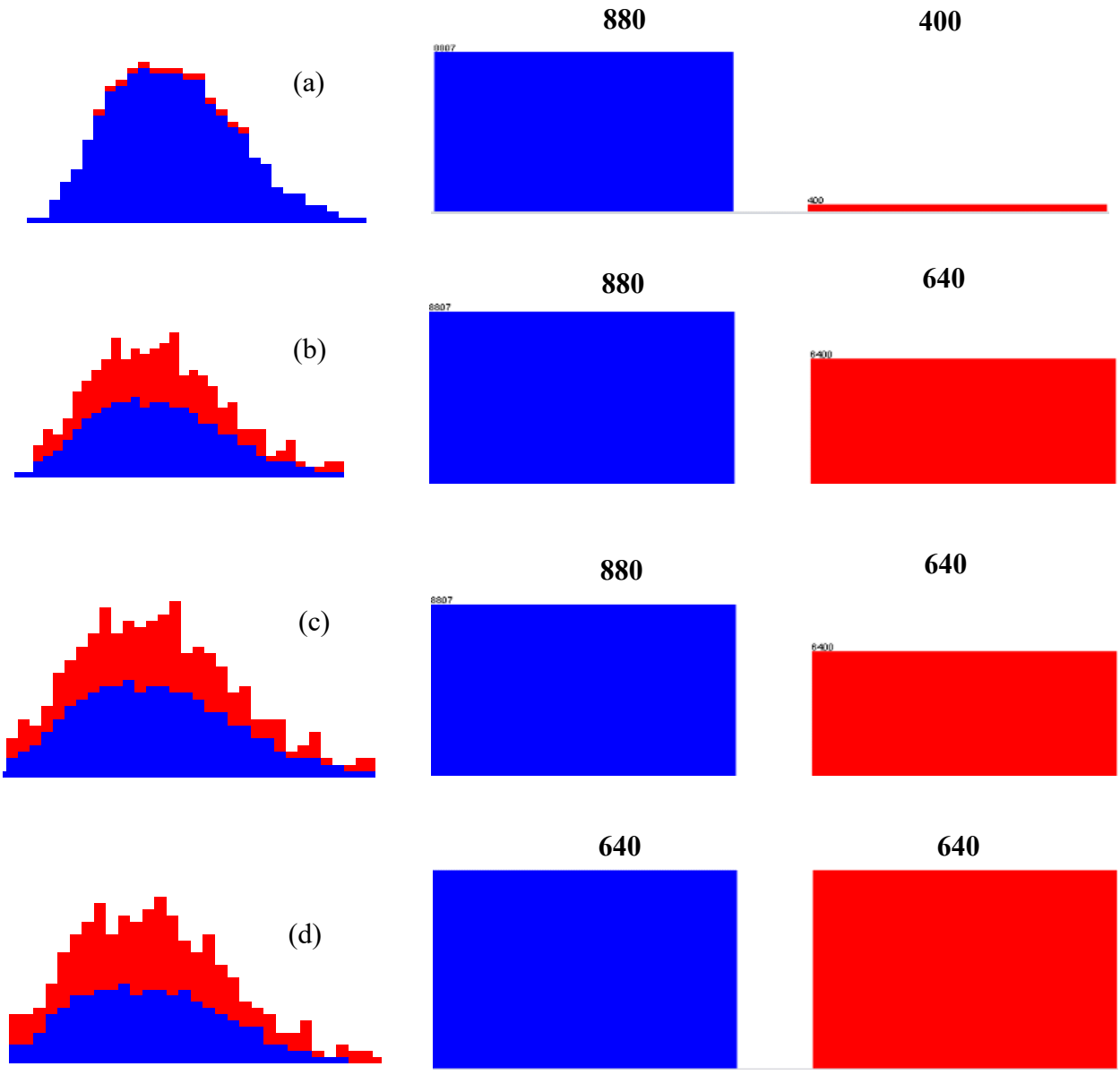
Appendix Figure C: White eggs MCCV classification accuracy and confusion matrix results on day 0 incubation

	Day 2							Day 3				
D2Br	M_F2	M_NF2	D2Br	M_F2	M_NF2		D3Br	M_F3	M_NF3	D3Br	M_F3	M_NF3
NF1		●	F12	●			NF1		●	F12		
NF2		●	F13				NF2		●	F13	●	
NF3		●	F14	●	●		NF3		●	F14	●	
NF4		●	F15	●	●		NF4	●	●	F15	●	
NF5		●	F16	●	●		NF5		●	F16		
NF6		●	F17	●	●		NF6	●	●	F17	●	
NF7		●	F18	●	●		NF7	●	●	F18	●	
NF8		●	F19	●	●		NF8	●	●	F19	●	
NF9		●	F20	●	●		NF9	●	●	F20	●	●
NF10		●	F21	●	●		NF10	●	●	F21	●	●
NF11		●	F22	●	●		NF11	●	●	F22	●	●
NF12		●	F23	●	●		NF12	●	●	F23	●	●
NF13		●	F24	●	●		NF13		●	F24	●	●
NF14	●	●	F25	●	●		NF14	●	●	F25	●	●
NF15	●	●	F26	●	●		NF15	●	●	F26	●	●
NF16	●	●	F27	●	●		NF16	●	●	F27	●	●
NF17		●	F28	●	●		NF17		●	F28	●	●
NF18		●	F29	●	●		NF18	●	●	F29	●	●
NF19		●	F30	●	●		NF19	●	●	F30	●	●
NF20		●	F31	●	●		NF20	●	●	F31	●	●
NF21		●	F32	●	●		NF21	●	●	F32	●	●
NF22		●	F33	●	●		NF22		●	F33	●	●
NF23		●	F34	●	●		NF23	●	●	F34	●	●
F1	●	●	F35	●	●		F1	●	●	F35	●	●
F2	●	●	F36	●	●		F2	●	●	F36	●	●
F3	●		F37	●	●		F3	●	●	F37	●	●
F4	●	●	F38	●	●		F4	●	●	F38	●	●
F5	●	●	F39	●	●		F5	●	●	F39	●	●
F6	●		F40	●	●		F6	●	●	F40	●	●
F7	●		F41	●	●		F7	●	●	F41	●	●
F8	●	●	F42	●	●		F8	●		F42	●	●
F9	●	●	F43	●	●		F9	●		F43	●	●
F10	●	●					F10	●	●			
F11	●	●					F11	●				

Appendix Figure D1: Typical SIMCA classification table for white eggs on different days of incubation

Day 2							Day 3					
D2Wh	M_F2	M_NF2	D2Wh	M_F2	M_NF2		D3Wh	M_F3	M_NF3	D3Wh	M_F3	M_NF3
NF1		•	F13	•	•		NF1		•	F13	•	•
NF2		•	F14	•	•		NF2		•	F14	•	•
NF3		•	F15	•	•		NF3		•	F15	•	•
NF4		•	F16	•	•		NF4		•	F16	•	•
NF5		•	F17	•	•		NF5		•	F17	•	•
NF6		•	F18	•	•		NF6		•	F18	•	•
NF7		•	F19	•	•		NF7		•	F19	•	•
NF8		•	F20	•	•		NF8		•	F20	•	•
NF9		•	F21	•	•		NF9		•	F21	•	•
NF10	•	•	F22	•	•		NF10		•	F22	•	•
NF11	•	•	F23	•	•		NF11		•	F23	•	•
NF12	•	•	F24	•	•		NF12		•	F24	•	•
NF13	•	•	F25	•	•		NF13		•	F25	•	•
NF14		•	F26	•	•		NF14		•	F26	•	•
NF15		•	F27	•	•		NF15	•	•	F27	•	•
NF16		•	F28	•	•		NF16	•	•	F28	•	•
NF17		•	F29	•	•		NF17	•	•	F29	•	•
NF18		•	F30	•	•		NF18	•	•	F30	•	•
NF19		•	F31	•	•		NF19	•	•	F31	•	•
NF20		•	F32	•	•		NF20		•	F32	•	•
NF21		•	F33	•	•		NF21		•	F33	•	•
F1	•	•	F34	•	•		F1	•	•	F34	•	•
F2	•	•	F35	•	•		F2	•	•	F35	•	•
F3	•	•	F36	•	•		F3	•	•	F36	•	•
F4	•	•	F37	•	•		F4	•	•	F37	•	•
F5	•	•	F38	•	•		F5	•	•	F38	•	•
F6	•	•	F39	•	•		F6	•	•	F39	•	•
F7	•	•	F40	•	•		F7	•	•	F40	•	•
F8	•	•	F41	•	•		F8	•	•	F41	•	•
F9	•	•	F42	•	•		F9	•	•	F42	•	•
F10	•	•	F43	•	•		F10	•	•			
F11	•	•	F44	•	•		F11	•	•			
F12	•	•					F12	•	•			

Appendix Figure D2: Typical SIMCA classification table for white eggs on different days of incubation



Appendix Figure E: Typical distribution profile (a), unsmote (b), smote (c) smote + randomization (d), smote + random under-sampling (RU)

REFERENCES

Abdel-Nour, Nicolas, Ngadi, Michael, Prasher, Shiv, & Karimi, Yousef. (2011). Prediction of egg freshness and albumen quality using visible/near infrared spectroscopy. *Food and Bioprocess Technology, 4*(5), 731-736.

Abdi, Hervé. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(1), 97-106.

Abdi, Hervé, & Williams, Lynne J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics, 2*(4), 433-459.

Aha, David W, Kibler, Dennis, & Albert, Marc K. (1991). Instance-based learning algorithms. *Machine learning, 6*(1), 37-66.

Aha, David W, & Kibler, Dennis F. (1989). *Noise-Tolerant Instance-Based Learning Algorithms.* Paper presented at the IJCAI.

Alejo, Roberto, García, Vicente, Sotoca, José Martínez, Mollineda, Ramón Alberto, & Sánchez, José Salvador. (2007). Improving the performance of the RBF neural networks trained with imbalanced samples *Computational and Ambient Intelligence* (pp. 162-169): Springer.

Alexandratos, Nikos, & Bruinsma, Jelle. (2012). World agriculture towards 2030/2050: the 2012 revision: ESA Working paper FAO, Rome.

Altman, Douglas G, & Bland, J Martin. (1994). Diagnostic tests 3: receiver operating characteristic plots. *BMJ: British Medical Journal, 309*(6948), 188.

Amigo, José Manuel, Babamoradi, Hamid, & Elcoroaristizabal, Saioa. (2015). Hyperspectral image analysis. A tutorial. *Analytica chimica acta, 896*, 34-51.

Amigo, José Manuel, Ravn, Carsten, Gallagher, Neal B, & Bro, Rasmus. (2009). A comparison of a common approach to partial least squares-discriminant analysis and classical least squares in hyperspectral imaging. *International journal of pharmaceutics, 373*(1-2), 179-182.

Andrew, NG. (2018). *Techinical Strategy for AI Engineers in the Era of Deep Learning. In: Machine Learning Yearning Draft Version*. Deeplearning.ai project.

Anton, Marc. (2007). Composition and structure of hen egg yolk *Bioactive egg compounds* (pp. 1-6): Springer.

Ariana, Diwan P, & Lu, Renfu. (2008). Detection of internal defect in pickling cucumbers using hyperspectral transmittance imaging. *Transactions of the ASABE, 51*(2), 705-713.

Ariana, Diwan P., & Lu, Renfu. (2010a). Hyperspectral imaging for defect detection of pickling cucumbers *Hyperspectral imaging for food quality analysis and control* (pp. 431-447): Elsevier.

Ariana, Diwan P., & Lu, Renfu. (2010b). Hyperspectral waveband selection for internal defect detection of pickling cucumbers and whole pickles. *Computers and Electronics in Agriculture, 74*(1), 137-144. doi: http://dx.doi.org/10.1016/j.compag.2010.07.008

Artís, Manuel, Ayuso, Mercedes, & Guillén, Montserrat. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance, 69*(3), 325-340.

Bakst, MR. (2010). Predicting Fertility: Section 2.: Visual and microscopic approaches for differentiating unfertilized germinal discs and early dead embryos from pre-incubated blastoderms. In M. R. Bakst & J. A. Long (Eds.), *Techniques for Semen Evaluation,*

Semen Storage, and Fertility Determination (2nd ed., pp. 79-85): The Midwest Poultry Federation, Buffalo, Minnesota.
Ballabio, D, & Todeschini, R. (2009). Multivariate classification for qualitative analysis.
Bamelis, F, Tona, K, De Baerdemaeker, J, & Decuypere, E. (2002). Detection of early embryonic development in chicken eggs using visible light transmission. *British poultry science, 43*(2), 204-212.
Bamelis, F., Kemps, B., Mertens, K., Tona, K., De Ketelaere, B., Decuypere, E., & De Baerdemaeker, J. (2004). Non-destructive measurements on eggs during incubation. *Avian and Poultry Biology Reviews, 15*(3-4), 150-159. doi: Doi 10.3184/147020604783637994
Barandela, Ricardo, Valdovinos, Rosa Maria, & Sánchez, José Salvador. (2003). New applications of ensembles of classifiers. *Pattern Analysis & Applications, 6*(3), 245-256.
Barker, Matthew, & Rayens, William. (2003). Partial least squares for discrimination. *Journal of chemometrics, 17*(3), 166-173.
Batista, Gustavo EAPA, Prati, Ronaldo C, & Monard, Maria Carolina. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter, 6*(1), 20-29.
Batuwita, Rukshan, & Palade, Vasile. (2010). *Efficient resampling methods for training support vector machines with imbalanced datasets.* Paper presented at the The 2010 International Joint Conference on Neural Networks (IJCNN).
Belitz, Hans-Dieter, Grosch, Werner, & Schieberle, Peter. (2009). Eggs. *Food chemistry*, 546-562.
Bennett, Kristin P, & Campbell, Colin. (2000). Support vector machines: hype or hallelujah? *Acm Sigkdd Explorations Newsletter, 2*(2), 1-13.
Bijlsma, Sabina, Bobeldijk, Ivana, Verheij, Elwin R, Ramaker, Raymond, Kochhar, Sunil, Macdonald, Ian A, . . . Smilde, Age K. (2006). Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Analytical chemistry, 78*(2), 567-574.
Bodkin, Andrew. (1997). Hyperspectral imaging at the speed of light. *analysis, 2561*, 80-97.
Bodkin, Andrew, Sheinis, Andrew, & Norton, Adam. (2005). Hyperspectral imaging systems: Google Patents.
Bouckaert, Remco R, Frank, Eibe, Hall, Mark, Kirkby, Richard, Reutemann, Peter, Seewald, Alex, & Scuse, David. (2010). WEKA manual for version 3-7-3. *University of Waikato, New Zealand*, 216.
Bradley, Andrew P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145-1159.
Brauns, Eric B, & Dyer, R Brian. (2006). Fourier transform hyperspectral visible imaging and the nondestructive analysis of potentially fraudulent documents. *Applied spectroscopy, 60*(8), 833-840.
Breiman, Leo. (1996). Stacked regressions. *Machine learning, 24*(1), 49-64.
Breiman, Leo. (2001). Random forests. *Machine learning, 45*(1), 5-32.
Breiman, Leo, Friedman, Jerome, Stone, Charles J, & Olshen, Richard A. (1984). *Classification and regression trees*: CRC press.
Briandet, Romain, Kemsley, E Katherine, & Wilson, Reginald H. (1996). Discrimination of Arabica and Robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. *Journal of agricultural and food chemistry, 44*(1), 170-174.

Bro, Rasmus, & Smilde, Age K. (2014). Principal component analysis. *Analytical Methods, 6*(9), 2812-2831.

Brown, Christopher D, & Davis, Herbert T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems, 80*(1), 24-38.

Bruce, BC, Carragher, BO, Damon, BM, Dawson, MJ, Eurell, JA, Gregory, CD, . . . Morris, HD. (1997). Chickscope: an interactive MRI classroom curriculum innovation for K-12. *Computers & Education, 29*(2-3), 73-87.

Bunkhumpornpat, Chumphol, Sinapiromsaran, Krung, & Lursinsap, Chidchanok. (2009). *Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem.* Paper presented at the Pacific-Asia conference on knowledge discovery and data mining.

Bunkhumpornpat, Chumphol, Sinapiromsaran, Krung, & Lursinsap, Chidchanok. (2012). DBSMOTE: density-based synthetic minority over-sampling technique. *Applied Intelligence, 36*(3), 664-684.

Bylesjö, Max, Rantalainen, Mattias, Cloarec, Olivier, Nicholson, Jeremy K., Holmes, Elaine, & Trygg, Johan. (2006). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics, 20*(8-10), 341-351. doi: 10.1002/cem.1006

Calvini, Rosalba, Ulrici, Alessandro, & Amigo, Jose Manuel. (2015). Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging. *Chemometrics and Intelligent Laboratory Systems, 146*, 503-511.

Camo, ASA. (2018). Unscrambler Users Guide and help documents, ver. 10.5.1. *Programme Package for Multivariate Calibration. Trondheim, Norway*.

Cao, Peng, Liu, Xiaoli, Bao, Hang, Yang, Jinzhu, & Zhao, Dazhe. (2015). Restricted Boltzmann machines based oversampling and semi-supervised learning for false positive reduction in breast CAD. *Bio-medical materials and engineering, 26*(s1), S1541-S1547.

Cassel, Claes, Hackl, Peter, & Westlund, Anders H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of applied statistics, 26*(4), 435-446.

Chao, Kuanglin, Yang, Chun-Chieh, Chen, Yud-Ren, Kim, Moon S, & Chan, Diane E. (2007). Fast line-scan imaging system for broiler carcass inspection. *Sensing and Instrumentation for Food Quality and Safety, 1*(2), 62-71.

Chawla, Nitesh V. (2009). Data mining for imbalanced datasets: An overview *Data mining and knowledge discovery handbook* (pp. 875-886): Springer.

Chawla, Nitesh V, Bowyer, Kevin W, Hall, Lawrence O, & Kegelmeyer, W Philip. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*(1), 321-357.

Chawla, Nitesh V, Cieslak, David A, Hall, Lawrence O, & Joshi, Ajay. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery, 17*(2), 225-252.

Chawla, Nitesh V, Japkowicz, Nathalie, & Kotcz, Aleksander. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter, 6*(1), 1-6.

Chawla, Nitesh V, Lazarevic, Aleksandar, Hall, Lawrence O, & Bowyer, Kevin W. (2003). *SMOTEBoost: Improving prediction of the minority class in boosting.* Paper presented at the European conference on principles of data mining and knowledge discovery.

Chen, Meng, Zhang, Lurui, & Xu, Huirong. (2015). On-line detection of blood spot introduced into brown-shell eggs using visible absorbance spectroscopy. *Biosystems Engineering, 131*, 95-101.

Chen, YR. (1993). *Classifying diseased poultry carcasses by visible and near-IR reflectance spectroscopy.* Paper presented at the Applications in Optical Science and Engineering.

Chen, YR, Chao, Kuanglin, & Kim, Moon S. (2002). Machine vision technology for agricultural applications. *Computers and electronics in Agriculture, 36*(2), 173-191.

Chen, YR, Park, B, Huffman, RW, & Nguyen, M. (1998). CLASSIFICATION OF ON-LINE POULTRY CARCASSES WITH BACKPROPAGATION NEURAL NETWORKS1. *Journal of Food Process Engineering, 21*(1), 33-48.

Chun, Hyonho, & Keleş, Sündüz. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(1), 3-25.

Chung, Dongjun, & Keles, Sunduz. (2010). Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology, 9*(1).

Clark, Matthew, & Cramer, Richard D. (1993). The probability of chance correlation using partial least squares (PLS). *Quantitative Structure-Activity Relationships, 12*(2), 137-145.

Cohen, Gilles, Hilario, Mélanie, Sax, Hugo, Hugonnet, Stéphane, & Geissbuhler, Antoine. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial intelligence in medicine, 37*(1), 7-18.

Coutts, JA, & Wilson, GC. (2007). Formation of the egg. *Optimum egg quality-A Practical Approach*.

Cover, Thomas, & Hart, Peter. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13*(1), 21-27.

Cox, Trevor F, & Cox, Michael AA. (2000). *Multidimensional scaling*: CRC Press.

Cristianini, Nello, & Shawe-Taylor, John. (2000). *An introduction to support vector machines and other kernel-based learning methods*: Cambridge university press.

Dale, Laura M, Thewis, André, Boudry, Christelle, Rotar, Ioan, Dardenne, Pierre, Baeten, Vincent, & Pierna, Juan A Fernández. (2013). Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: a review. *Applied Spectroscopy Reviews, 48*(2), 142-159.

Das, K, & Evans, MD. (1992a). Detecting fertility of hatching eggs using machine vision I. Histogram characterization method. *Transactions of the ASAE, 35*(4), 1335-1341.

Das, K, & Evans, MD. (1992b). Detecting fertility of hatching eggs using machine vision. II. Neural network classifiers. *Transactions of the ASAE (USA)*.

Davis, Jesse, & Goadrich, Mark. (2006). *The relationship between Precision-Recall and ROC curves.* Paper presented at the Proceedings of the 23rd international conference on Machine learning.

Del Fiore, A, Reverberi, M, Ricelli, A, Pinzari, F, Serranti, S, Fabbri, AA, . . . Fanelli, C. (2010). Early detection of toxigenic fungi on maize by hyperspectral imaging analysis. *International journal of food microbiology, 144*(1), 64-71.

Dong, Guozhu, Zhang, Xiuzhen, Wong, Limsoon, & Li, Jinyan. (1999). *CAEP: Classification by aggregating emerging patterns.* Paper presented at the Discovery Science.

Du, Cheng-Jin, & Sun, Da-Wen. (2004). Recent developments in the applications of image processing techniques for food quality evaluation. *Trends in food science & technology, 15*(5), 230-249.

Du, Cheng-Jin, & Sun, Da-Wen. (2006). Learning techniques used in computer vision for food quality evaluation: a review. *Journal of food engineering, 72*(1), 39-55.

Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R., & Griffin, J. L. (2011). Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev, 40*(1), 387-426. doi: 10.1039/b906712b

Dunn, Warwick B, Wilson, Ian D, Nicholls, Andrew W, & Broadhurst, David. (2012). The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans.

Eavis, Todd, & Japkowicz, Nathalie. (2000). A recognition-based alternative to discrimination-based multi-layer perceptrons *Advances in Artificial Intelligence* (pp. 280-292): Springer.

Efron, Bradley, & Tibshirani, Robert. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association, 92*(438), 548-560.

Egan, JP. (1975). Signal detection theory and ROC analysis. Series in Cognition and Perception.: Academic Press, New York.

Elkan, Charles. (2001). *The foundations of cost-sensitive learning.* Paper presented at the International joint conference on artificial intelligence.

ElMasry, Gamal, Kamruzzaman, Mohammed, Sun, Da-Wen, & Allen, Paul. (2012). Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review. *Critical reviews in food science and nutrition, 52*(11), 999-1023.

ElMasry, Gamal M, & Nakauchi, Shigeki. (2016). Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality–A comprehensive review. *Biosystems Engineering, 142*, 53-82.

ElMasry, Gamal, & Sun, Da-Wen. (2010a). Meat quality assessment using a hyperspectral imaging system. In D.-W. Sun (Ed.), *Hyperspectral imaging for food quality analysis and control* (pp. 175-240). San Diego: Academic Press.

ElMasry, Gamal, & Sun, Da-Wen. (2010b). Principles of hyperspectral imaging technology. In D.-W. Sun (Ed.), *Hyperspectral imaging for food quality analysis and control* (pp. 3-43). San Diego: Academic Press.

ElMasry, Gamal, Sun, Da-Wen, & Allen, Paul. (2013). Chemical-free assessment and mapping of major constituents in beef using hyperspectral imaging. *Journal of Food Engineering, 117*(2), 235-246.

Eriksson, Lennart, Byrne, T, Johansson, E, Trygg, Johan, & Vikström, C. (2013). *Multi-and megavariate data analysis basic principles and applications*: Umetrics Academy.

Ernst, RA, Bradley, FA, Abbott, UK, & Craig, RM. (2004). Egg candling and breakout analysis for hatchery quality assurance and analysis of poor hatches. *University of California, Division of Agriculture and Natural Resources. pub, 8134*.

Esquerre, C, Gowen, AA, Burger, J, Downey, G, & O'Donnell, CP. (2012). Suppressing sample morphology effects in near infrared spectral imaging using chemometric data pre-treatments. *Chemometrics and Intelligent Laboratory Systems, 117*, 129-137.

Fawcett, Tom. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Fawcett, Tom, & Provost, Foster. (1997). Adaptive fraud detection. *Data mining and knowledge discovery, 1*(3), 291-316.

Feng, Aimin, & Chen, Bin. (2008). *Learning the boundary of One-Class-Classifier globally and locally.* Paper presented at the Cybernetics and Intelligent Systems, 2008 IEEE Conference on.

Fernández, Alberto, del Jesus, María José, & Herrera, Francisco. (2010). On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. *Information Sciences, 180*(8), 1268-1291.

Fernández, Alberto, García, Salvador, & Herrera, Francisco. (2011). Addressing the classification with imbalanced data: open problems and new challenges on class distribution *Hybrid Artificial Intelligent Systems* (pp. 1-10): Springer.

Ferri, FJ, Pudil, Pavel, Hatef, Mohamad, & Kittler, Josef. (1994). Comparative study of techniques for large-scale feature selection *Machine Intelligence and Pattern Recognition* (Vol. 16, pp. 403-413): Elsevier.

Filzmoser, Peter, Gschwandtner, Moritz, & Todorov, Valentin. (2012). Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics, 26*(3-4), 42-51.

Filzmoser, Peter, Liebmann, Bettina, & Varmuza, Kurt. (2009). Repeated double cross validation. *Journal of Chemometrics, 23*(4), 160-171.

Forman, George. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research, 3*(Mar), 1289-1305.

François, D. (2006). Binary classification performances measure cheat sheet. *Journal of Machine Learning Research, 7*, 1-30.

Frank, Eibe, Hall, Mark, Trigg, Len, Holmes, Geoffrey, & Witten, Ian H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics, 20*(15), 2479-2481.

Galar, Mikel, Fernandez, Alberto, Barrenechea, Edurne, Bustince, Humberto, & Herrera, Francisco. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 463-484.

García, Salvador, & Herrera, Francisco. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary computation, 17*(3), 275-306.

Ghaoui, Laurent E, Jordan, Michael I, & Lanckriet, Gert R. (2003). *Robust novelty detection with single-class MPM.* Paper presented at the Advances in neural information processing systems.

Giangiacomo, Roberto, & Dull, Gerald. (1986). Near infrared spectrophotometric determination of individual sugars in aqueous mixtures. *Journal of Food Science, 51*(3), 679-683.

Good, Phillip I. (2011a). *Analyzing the large number of variables in biomedical and satellite imagery*: John Wiley & Sons.

Good, Phillip I. (2011b). *Permutation tests*: Wiley Online Library.

Gottfries, Johan, Blennow, Kaj, Wallin, Anders, & Gottfries, CG. (1995). Diagnosis of dementias using partial least squares discriminant analysis. *Dementia and Geriatric Cognitive Disorders, 6*(2), 83-88.

Guyon, Isabelle, & Elisseeff, André. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.

Hackstadt, Amber J, & Hess, Ann M. (2009). Filtering for increased power for microarray data analysis. *BMC bioinformatics, 10*(1), 11.

Hajmeer, M, & Basheer, I. (2003). Comparison of logistic regression and neural network-based classifiers for bacterial growth. *Food Microbiology, 20*(1), 43-55.
Hall, Mark Andrew. (1999). Correlation-based feature selection for machine learning.
Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, & Witten, Ian H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.
Halls, Amy, & Shur-Gain. (2014). Egg formation and eggshell quality in layers. *Nutrifax Nutrition News and Information Update*, 1-5.
Han, Hui, Wang, Wen-Yuan, & Mao, Bing-Huan. (2005). *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning.* Paper presented at the International Conference on Intelligent Computing.
Hanley, James A, & McNeil, Barbara J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.
He, Haibo, Bai, Yang, Garcia, Edwardo A, & Li, Shutao. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* Paper presented at the Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on.
He, Haibo, & Garcia, Edwardo A. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on, 21*(9), 1263-1284.
Hernández-Hernández, A, Vidal, ML, Gómez-Morales, J, Rodríguez-Navarro, AB, Labas, V, Gautron, J, . . . Ruiz, JM García. (2008). Influence of eggshell matrix proteins on the precipitation of calcium carbonate (CaCO 3). *Journal of Crystal Growth, 310*(7), 1754-1759.
Ho, Tin Kam. (1995). *Random decision forests.* Paper presented at the Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on.
Höskuldsson, Agnar. (1988). PLS regression methods. *Journal of chemometrics, 2*(3), 211-228.
Höskuldsson, Agnar. (1996). *Prediction methods in science and technology. 1. Basic theory* (Vol. 1): Thor Publ.
Hosmer, DW, & Lemeshow, S. (2000). Applied logistic regression, John Wiley&Sons New York. *Applied logistic regression. 2nd ed. John Wiley & Sons, New York.*, -.
Hosmer Jr, David W, Lemeshow, Stanley, & Sturdivant, Rodney X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.
Hotelling, Harold. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology, 24*(6), 417.
Hukerikar, Saumil, Tumma, Ashwin, Nikam, Akshay, & Attar, Vahida. (2011). *SkewBoost: An algorithm for classifying imbalanced datasets.* Paper presented at the Computer and Communication Technology (ICCCT), 2011 2nd International Conference on.
Iizuka, K, & Aishima, T. (1997). Soy sauce classification by geographic region based on NIR spectra and chemometrics pattern recognition. *Journal of food science, 62*(1), 101-104.
Japkowicz, Nathalie. (2000). *Learning from imbalanced data sets: a comparison of various strategies.* Paper presented at the AAAI workshop on learning from imbalanced data sets.
Japkowicz, Nathalie. (2001). Concept-learning in the presence of between-class and within-class imbalances *Advances in Artificial Intelligence* (pp. 67-77): Springer.
Japkowicz, Nathalie, Myers, Catherine, & Gluck, Mark. (1995). *A novelty detection approach to classification.* Paper presented at the IJCAI.

Japkowicz, Nathalie, & Stephen, Shaju. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis, 6*(5), 429-449.
Jason, B. (2016). *Machine Learning Mastery With Weka: Analyse data, develop models and work through projects Machine Learning Mastery* (pp. 1-248).
Jestel, Nancy L, Shaver, Jeremy M, & Morris, Michael D. (1998). Hyperspectral Raman line imaging of an aluminosilicate glass. *Applied spectroscopy, 52*(1), 64-69.
Jindal, Nitin, & Liu, Bing. (2007). *Review spam detection.* Paper presented at the Proceedings of the 16th international conference on World Wide Web.
Jo, Taeho, & Japkowicz, Nathalie. (2004). Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter, 6*(1), 40-49.
Johnson, W Evan, Li, Cheng, & Rabinovic, Ariel. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127.
Kamruzzaman, Mohammed, ElMasry, Gamal, Sun, Da-Wen, & Allen, Paul. (2013). Non-destructive assessment of instrumental and sensory tenderness of lamb meat using NIR hyperspectral imaging. *Food Chemistry, 141*(1), 389-396.
Khan, Shehroz S., & Madden, Michael G. (2010). *A Survey of Recent Trends in One Class Classification*, Berlin, Heidelberg.
Kim, MS, Lefcourt, AM, Chao, K, Chen, YR, Kim, I, & Chan, DE. (2002). Multispectral detection of fecal contamination on apples based on hyperspectral imagery: Part I. Application of visible and near-infrared reflectance imaging. *TRANSACTIONS-AMERICAN SOCIETY OF AGRICULTURAL ENGINEERS, 45*(6), 2027-2038.
King'Ori, AM. (2011). Review of the factors that influence egg fertility and hatchability in poultry. *International Journal of Poultry Science, 10*(6), 483-492.
Klein, Marvin E, Aalderink, Bernard J, Padoan, Roberto, De Bruin, Gerrit, & Steemers, Ted AG. (2008). Quantitative hyperspectral reflectance imaging. *Sensors, 8*(9), 5576-5618.
Kleinbaum, DG, & Klein, M. (1994). Logistic regression: a self-learning text Springer-Verlag. *New York, NY*.
Kubat, Miroslav, Holte, Robert C, & Matwin, Stan. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine learning, 30*(2-3), 195-215.
Kubat, Miroslav, & Matwin, Stan. (1997). *Addressing the curse of imbalanced training sets: one-sided selection.* Paper presented at the Icml.
Kudo, Mineichi, & Sklansky, Jack. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern recognition, 33*(1), 25-41.
Kuhn, Max, & Johnson, Kjell. (2013). *Applied predictive modeling* (Vol. 26): Springer.
Kuhn, Max, & Johnson, Kjell. (2016). Remedies for severe class imbalance *Applied predictive modeling* (pp. 427): Springer.
Kuncheva, Ludmila I, & Rodríguez, Juan J. (2014). A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems, 38*(2), 259-275.
Kurt, Imran, Ture, Mevlut, & Kurum, A Turhan. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert systems with applications, 34*(1), 366-374.
Ladha, L, & Deepa, T. (2011). Feature selection methods and algorithms. *International journal on computer science and engineering, 3*(5), 1787-1797.
Latour, Mickey A, Meunier, Ryan, & Stewart, Jessica. (2014). Poultry: The process of egg formation. *Purdue University Cooperative Extension Service, West Lafayette IN, Issued in furtherance of the acts of May 8 and June 30, 1914.*

Laurikkala, Jorma. (2001). *Improving identification of difficult small classes by balancing class distribution.* Paper presented at the Conference on Artificial Intelligence in Medicine in Europe.

Lawrence, KC, Park, Bosoon, Heitschmidt, GW, Windham, WR, & Thai, CN. (2007). Evaluation of LED and tungsten-halogen lighting for fecal contaminant detection. *Applied engineering in agriculture, 23*(6), 811-818.

Lawrence, Kurt, Smith, Douglas, Windham, William, Heitschmidt, Gerald, & Park, Bosoon. (2006). Egg embryo development detection with hyperspectral imaging. *International Journal of Poultry Science, 5*(10), 964-969.

Lê Cao, Kim-Anh, Boitard, Simon, & Besse, Philippe. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics, 12*(1), 253.

Lê Cao, Kim-Anh, Bonnet, Agnès, & Gadat, Sébastien. (2009). Multiclass classification and gene selection with a stochastic algorithm. *Computational Statistics & Data Analysis, 53*(10), 3601-3615.

Lê Cao, Kim-Anh, Martin, Pascal GP, Robert-Granié, Christèle, & Besse, Philippe. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics, 10*(1), 34.

Lê Cao, Kim-Anh, Rossouw, Debra, Robert-Granié, Christele, & Besse, Philippe. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology, 7*(1).

Lewis, David, & Gale, William. (1994). Training text classifiers by uncertainty sampling.

Lewis, E Neil, Dubois, Janie, Kidder, Linda H, & Haber, Kenneth S. (2007). Near infrared chemical imaging: beyond the pictures. *Techniques and Applications of Hyperspectral Image Analysis*, 335-361.

Li, Jinyan, Dong, Guozhu, Ramamohanarao, Kotagiri, & Wong, Limsoon. (2004). Deeps: A new instance-based lazy discovery and classification system. *Machine Learning, 54*(2), 99-124.

Liao, T Warren. (2008). Classification of weld flaws with imbalanced class data. *Expert Systems with Applications, 35*(3), 1041-1052.

Liaw, Andy, & Wiener, Matthew. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Liebmann, Bettina, Filzmoser, Peter, & Varmuza, Kurt. (2010). Robust and classical PLS regression compared. *Journal of Chemometrics, 24*(3-4), 111-120.

Ling, Charles X, Yang, Qiang, Wang, Jianning, & Zhang, Shichao. (2004). *Decision trees with minimal costs.* Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Liu, D, Sun, Da-Wen, & Zeng, Xin-An. (2014). Recent advances in wavelength selection techniques for hyperspectral image processing in the food industry. *Food and Bioprocess Technology, 7*(2), 307-323.

Liu, L., & Ngadi, M. O. (2013). Detecting Fertility and Early Embryo Development of Chicken Eggs Using Near-Infrared Hyperspectral Imaging. *Food and Bioprocess Technology, 6*(9), 2503-2513. doi: 10.1007/s11947-012-0933-3

Liu, X, Wu, J, & Zhou, Z. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(2), 539-550.

Liu, Y, Chawla, Nitesh V, Harper, Mary P, Shriberg, Elizabeth, & Stolcke, Andreas. (2006). A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language, 20*(4), 468-494.

Longadge, Rushi, & Dongre, Snehalata. (2013). Class Imbalance Problem in Data Mining Review. *arXiv preprint arXiv:1305.1707*.

López, Victoria, Fernández, Alberto, García, Salvador, Palade, Vasile, & Herrera, Francisco. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141.

Mahadevan, Sankar, Shah, Sirish L, Marrie, Thomas J, & Slupsky, Carolyn M. (2008). Analysis of metabolomic data using support vector machines. *Analytical Chemistry, 80*(19), 7562-7570.

Manevitz, Larry, & Yousef, Malik. (2002). One-class SVMs for document classification. *the Journal of machine Learning research, 2*, 139-154.

Manevitz, Larry, & Yousef, Malik. (2007). One-class document classification via neural networks. *Neurocomputing, 70*(7), 1466-1481.

Mani, Inderjeet, & Zhang, I. . (2003). *kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction.* Paper presented at the Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets, Washington, DC.

Mehl, Patrick M, Chen, Yud-Ren, Kim, Moon S, & Chan, Diane E. (2004). Development of hyperspectral imaging technique for the detection of apple surface defects and contaminations. *Journal of Food Engineering, 61*(1), 67-81.

Meyer, David. (2017). Support vector machines: The interface to libsvm in package e1071.

Meyer, David, & Wien, FH Technikum. (2001). Support vector machines. *R News, 1*(3), 23-26.

Mottet, A, & Tempio, G. (2017). Global poultry production: current state and future outlook and challenges. *World's Poultry Science Journal, 73*(2), 245-256.

Mueller-Mach, Regina, Mueller, Gerd O, Krames, Michael R, & Trottier, Troy. (2002). High-power phosphor-converted light-emitting diodes based on III-nitrides. *Selected Topics in Quantum Electronics, IEEE Journal of, 8*(2), 339-345.

Munroe, Daniel T, & Madden, Michael G. (2005). Multi-class and single-class classification approaches to vehicle model recognition from images. *Proc. AICS*, 1-11.

Naik, Ganesh R, & Kumar, Dinesh Kant. (2011). An overview of independent component analysis and its applications. *Informatica: An International Journal of Computing and Informatics, 35*(1), 63-81.

Nakariyakul, Songyot. (2007). *Feature selection algorithms for anomaly detection in hyperspectral data.* Carnegie Mellon University.

NASS, USDA. (2006). Statistical Highlights of US Agriculture, 2005-2006.

Nguyen, C, & Ho, T. (2005). An imbalanced data rule learner *Knowledge Discovery in Databases: PKDD 2005* (pp. 617-624): Springer.

Nguyen, G, Bouzerdoum, A, & Phung, S. (2009). Learning pattern classification tasks with imbalanced data sets. In P. Yin (Ed.), *Pattern recognition* (pp. 193-208). Vukovar, Croatia: INTECH Open Access Publisher.

Noh, Hyun Kwon, & Lu, Renfu. (2007). Hyperspectral laser-induced fluorescence imaging for assessing apple fruit quality. *Postharvest Biology and Technology, 43*(2), 193-201.

Norris, KH. (1996). History of NIR. *Journal of Near Infrared Spectroscopy, 4*, 31-38.

Okubo, T, Akachi, S, & Hatta, H. (1997). Structure of hen eggs and physiology of egg laying. In T. Yamamoto, L. R. Juneja, H. Hatta & M. Kim (Eds.), *Hen Eggs: Basic and Applied Science*: CRC press, New York.

Ortiz, M Cruz, Sarabia, Luis A, Symington, Charles, Santamaría, Fernando, & Íñiguez, Montserrat. (1996). Analysis of ageing and typification of vintage ports by partial least squares and soft independent modelling class analogy. *Analyst, 121*(8), 1009-1013.

Osborne, Brian G, Fearn, Thomas, & Hindle, Peter H. (1993). *Practical NIR spectroscopy with applications in food and beverage analysis*: Longman scientific and technical.

Özdamar, K. (2004). Paket programlarla istatistiksel veri analizi. Kaan kitabevi, Eskişehir'den aktaran Akbulut, Y. *Sosyal Bilimlerde SPSS Uygulamaları*, 632-633.

Partridge, Matthew, & Calvo, Rafael A. (1998). Fast dimensionality reduction and simple PCA. *Intelligent Data Analysis, 2*(3), 203-214.

Pasquini, Celio. (2003). Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society, 14*(2), 198-219.

Pearson, Karl. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559-572.

Pérez-Enciso, Miguel, & Tenenhaus, Michel. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human genetics, 112*(5-6), 581-592.

Phoungphol, Piyaphol. (2013). *A classification framework for imbalanced data.* (PhD), Georgia State University.

Phua, Clifton, Alahakoon, Damminda, & Lee, Vincent. (2004). Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter, 6*(1), 50-59.

Picard, Richard R, & Cook, R Dennis. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79*(387), 575-583.

Polikar, Robi. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine, 6*(3), 21-45.

Potter, Clinton S, Weber, Daniel E, Sinn-Hanlon, Janet, Fossum, Barbara Mason, Bruce, C Bertram, & Thakkar, Umesh. (1998a). Chickscope 1.5: Explore Embryology Day 1- The Journey Begins: An MRI's View.

Potter, Clinton S, Weber, Daniel E, Sinn-Hanlon, Janet, Fossum, Barbara Mason, Bruce, C Bertram, & Thakkar, Umesh. (1998b). Chickscope 1.5: Explore Embryology Day 1-The Journey Begins: Meiosis and Mitosis.

Prachuabsupakij, Wanthanee. (2015). *CLUS: A new hybrid sampling classification for imbalanced data.* Paper presented at the Computer Science and Software Engineering (JCSSE), 2015 12th International Joint Conference, Thailand.

Provost, Foster J, & Fawcett, Tom. (1997). *Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions.* Paper presented at the KDD.

Provost, Foster J, Fawcett, Tom, & Kohavi, Ron. (1998). *The case against accuracy estimation for comparing induction algorithms.* Paper presented at the ICML.

Pudil, Pavel, Novovičová, Jana, & Kittler, Josef. (1994). Floating search methods in feature selection. *Pattern recognition letters, 15*(11), 1119-1125.

Qin, J. (2010). Hyperspectral Imaging Instruments. In D.-W. Sun (Ed.), *Hyperspectral imaging for food quality analysis and control* (pp. 129-172). San Diego, California, USA: Elsevier.

Qin, J, & Lu, Renfu. (2005). Detection of pits in tart cherries by hyperspectral transmission imaging. *Transactions of the ASAE, 48*(5), 1963-1970.
Qin, Jianwei, Chao, Kuanglin, Kim, Moon S, Lu, Renfu, & Burks, Thomas F. (2013). Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering, 118*(2), 157-171.
Raghavan, Vijay, Bollmann, Peter, & Jung, Gwang S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS), 7*(3), 205-229.
Rahman, M Mostafizur, & Davis, DN. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing, 3*(2), 224.
Raskutti, Bhavani, & Kowalczyk, Adam. (2004). Extreme re-balancing for SVMs: a case study. *ACM Sigkdd Explorations Newsletter, 6*(1), 60-69.
Renard, Nadine, Bourennane, Salah, & Blanc-Talon, Jacques. (2008). Denoising and dimensionality reduction using multilinear tools for hyperspectral images. *Geoscience and Remote Sensing Letters, IEEE, 5*(2), 138-142.
Rokach, Lior. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1-2), 1-39.
Romanoff, Alexis Lawrence, & Romanoff, Anastasia J. (1949). *The avian egg*: John Wiley & Sons, Inc. New York.
Rothman, KJ, & Greenland, Sander. (1998). Modern epidemiology. 2nd. *Philadelphia, PA, Lippencott-Raven Publishers*.
Saeys, Yvan, Inza, Iñaki, & Larrañaga, Pedro. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507-2517.
Sahin, Serpil, & Sumnu, Servet Gülüm. (2006). *Physical properties of foods*: Springer Science & Business Media.
Saito, Takaya, & Rehmsmeier, Marc. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*(3), e0118432.
Sasaki, Yutaka. (2007). The truth of the F-measure. *Teach Tutor mater, 1*(5).
Schapire, Robert E. (1990). The strength of weak learnability. *Machine learning, 5*(2), 197-227.
Schapire, Robert E. (2003). The boosting approach to machine learning: An overview *Nonlinear estimation and classification* (pp. 149-171): Springer.
Schölkopf, Bernhard, Burges, Christopher JC, & Smola, Alexander J. (1999). *Advances in kernel methods: support vector learning*: MIT press.
Schölkopf, Bernhard, Platt, John C, Shawe-Taylor, John, Smola, Alex J, & Williamson, Robert C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation, 13*(7), 1443-1471.
Schuermann, Juergen, & Doster, Wolfgang. (1984). A decision theoretic approach to hierarchical classifier design. *Pattern Recognition, 17*(3), 359-369.
Schumacher, Martin, Roßner, Reinhard, & Vach, Werner. (1996). Neural networks and logistic regression: Part I. *Computational Statistics & Data Analysis, 21*(6), 661-682.
Schweizer, Susan M, & Moura, José MF. (2001). Efficient detection in hyperspectral imagery. *Image Processing, IEEE Transactions on, 10*(4), 584-597.
Seiffert, Chris, Khoshgoftaar, Taghi M, Van Hulse, Jason, & Napolitano, Amri. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40*(1), 185-197.

Senthilkumar, T, Jayas, DS, White, NDG, Fields, PG, & Gräfenhan, T. (2016a). Detection of fungal infection and Ochratoxin A contamination in stored wheat using near-infrared hyperspectral imaging. *Journal of Stored Products Research, 65*, 30-39.

Senthilkumar, T, Singh, CB, Jayas, DS, & White, NDG. (2012). Detection of fungal infection in canola using near-infrared hyperspectral imaging. *Journal of Agricultural Engineering, 49*(1), 21-27.

Senthilkumar, Thiruppathi, Jayas, Digvir S, White, Noel DG, Fields, Paul G, & Gräfenhan, Tom. (2016b). Detection of fungal infection and Ochratoxin A contamination in stored barley using near-infrared hyperspectral imaging. *Biosystems Engineering, 147*, 162-173.

Shetty, Sujala D, Vadivel, S, & Vaghella, Sakshi. (2010). Weka based desktop data mining as web service. *World academy of science, engineering and technology*, 697-715.

Shlens, Jonathon. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Siesler, Heinz W, Ozaki, Yukihiro, Kawata, Satoshi, & Heise, H Michael. (2008). *Near-infrared spectroscopy: principles, instruments, applications*: John Wiley & Sons.

Smit, Suzanne, van Breemen, Mariëlle J, Hoefsloot, Huub CJ, Smilde, Age K, Aerts, Johannes MFG, & De Koster, Chris G. (2007). Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta, 592*(2), 210-217.

Smith, DP, Lawrence, KC, & Heitschmidt, GW. (2008). Fertility and embryo development of broiler hatching eggs evaluated with a hyperspectral imaging and predictive modeling system. *International journal of poultry science, 7*(10), 1001-1004.

Smith, DP, Mauldin, JM, Lawrence, KC, Park, Bosoon, & Heitschmidt, GW. (2005). *Detection of fertility and early development of hatching eggs with hyperspectral imaging.* Paper presented at the Proc. 11th European Symposium on the Quality of Eggs and Egg Products Netherlands: World's Poultry Science Association.

Smola, Alexander J. (2000). *Advances in large margin classifiers*: MIT press.

Sokolova, Marina, & Lapalme, Guy. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

Sone, Izumi, Olsen, Ragnar L, Sivertsen, Agnar H, Eilertsen, Guro, & Heia, Karsten. (2012). Classification of fresh Atlantic salmon (Salmo salar L.) fillets stored under different atmospheres by hyperspectral imaging. *Journal of Food Engineering, 109*(3), 482-489.

Spackman, Kent A. (1989). *Signal detection theory: Valuable tools for evaluating inductive learning.* Paper presented at the Proceedings of the sixth international workshop on Machine learning.

Spinosa, Eduardo J, & de Carvalho, André CPLF. (2005). Combining one-class classifiers for robust novelty detection in gene expression data *Advances in bioinformatics and computational biology* (pp. 54-64): Springer.

Stadelman, W J. (2017). Quality identification of shell eggs *Egg Science and Technology, Fourth Edition* (pp. 55-82): CRC Press.

Stefanowski, Jerzy, & Wilk, Szymon. (2008). *Selective pre-processing of imbalanced data for improving classification performance.* Paper presented at the International Conference on Data Warehousing and Knowledge Discovery.

Steiner, Gerald, Bartels, Thomas, Stelling, Allison, Krautwald-Junghanns, Maria-Elisabeth, Fuhrmann, Herbert, Sablinskas, Valdas, & Koch, Edmund. (2011). Gender determination of fertilized unincubated chicken eggs by infrared spectroscopic imaging. *Analytical and bioanalytical chemistry, 400*(9), 2775-2782.

Sturkie, Paul D. (1965). *Avian Physiology Second Edition*: Ithaca, New York, Cornell Univ. Press.
Subasi, Abdulhamit, & Ercelebi, Ergun. (2005). Classification of EEG signals using neural network and logistic regression. *Computer methods and programs in biomedicine, 78*(2), 87-99.
Suguna, N, & Thanushkodi, K. (2010). An improved k-nearest neighbor classification using genetic algorithm. *International Journal of Computer Science Issues, 7*(2), 18-21.
Sun, Da-Wen. (2010). *Hyperspectral imaging for food quality analysis and control*: Elsevier.
Sun, Y, Wong, A, & Kamel, M. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(04), 687-719.
Sun, Yanmin, Kamel, Mohamed S, Wong, Andrew KC, & Wang, Yang. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*(12), 3358-3378.
Swarbrick, B. (2012). Multivariate data analysis for dummies. *CAMO software special edition*, 1-43.
Swets, John A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.
Swets, John A, Dawes, Robyn M, & Monahan, John. (2000). Better decisions through science. *Scientific American, 283*(4), 82-87.
Szymańska, Ewa, Saccenti, Edoardo, Smilde, Age K, & Westerhuis, Johan A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics, 8*(1), 3-16.
Tang, Yuchun, Zhang, Yan-Qing, Chawla, Nitesh V, & Krasser, Sven. (2009). SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(1), 281-288.
Tapp, Henri S, & Kemsley, E Kate. (2009). Notes on the practical utility of OPLS. *TrAC Trends in Analytical Chemistry, 28*(11), 1322-1327.
Teck, Ching Chuen, Xiang, Li, Junhong, Zhou, Xiaoli, Li, Hong, Cao, & Woon, David. (2012). *Hybrid rebalancing approach to handle imbalanced dataset for fault diagnosis in manufacturing systems.* Paper presented at the Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on.
Tibshirani, Robert. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
Tibshirani, Robert. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 73*(3), 273-282.
Tobias, Randall D. (1995). *An introduction to partial least squares regression.* Paper presented at the Proc. Ann. SAS Users Group Int. Conf., 20th, Orlando, FL.
Tomek, Ivan. (1976). Two modifications of CNN. *IEEE Trans. Systems, Man and Cybernetics, 6*, 769-772.
Trygg, Johan. (2002). O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of chemometrics, 16*(6), 283-293.
Trygg, Johan, & Wold, Svante. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of chemometrics, 16*(3), 119-128.
Vach, Werner, Roßner, Reinhard, & Schumacher, Martin. (1996). Neural networks and logistic regression: Part II. *Computational Statistics & Data Analysis, 21*, 683-701.

Van Hulse, Jason, Khoshgoftaar, Taghi M, & Napolitano, Amri. (2009). *An empirical comparison of repetitive undersampling techniques.* Paper presented at the Information Reuse & Integration, 2009. IRI'09. IEEE International Conference on.

Vapnik, Vladimir Naumovich, & Vapnik, Vlamimir. (1998). *Statistical learning theory* (Vol. 1): Wiley New York.

Vidal, Maider, & Amigo, José Manuel. (2012). Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems, 117*, 138-148.

Wabuyele, Musundi B, Yan, Fei, Griffin, Guy D, & Vo-Dinh, Tuan. (2005). Hyperspectral surface-enhanced Raman imaging of labeled silver nanoparticles in single cells. *Review of scientific instruments, 76*(6), 063710.

Wakeling, Ian N, & Morris, Jeff J. (1993). A test of significance for partial least squares regression. *Journal of Chemometrics, 7*(4), 291-304.

Wang, Ning, & ElMasry, Gamal. (2010). Bruise detection of apples using hyperspectral imaging *Hyperspectral imaging for food quality analysis and control* (pp. 295-320): Elsevier.

Wang, Shuo, & Yao, Xin. (2013). Relationships between diversity of classification ensembles and single-class performance measures. *IEEE Transactions on Knowledge and Data Engineering, 25*(1), 206-219.

Wasikowski, Mike, & Chen, Xue-wen. (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on knowledge and data engineering, 22*(10), 1388-1400.

Weiss, Gary M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter, 6*(1), 7-19.

Westerhuis, Johan A, Hoefsloot, Huub CJ, Smit, Suzanne, Vis, Daniel J, Smilde, Age K, van Velzen, Ewoud JJ, . . . van Dorsten, Ferdi A. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*(1), 81-89.

Westerhuis, Johan A, van Velzen, Ewoud JJ, Hoefsloot, Huub CJ, & Smilde, Age K. (2010). Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics, 6*(1), 119-128.

Weston, Jason, Mukherjee, Sayan, Chapelle, Olivier, Pontil, Massimiliano, Poggio, Tomaso, & Vapnik, Vladimir. (2001). *Feature selection for SVMs.* Paper presented at the Advances in neural information processing systems.

Williams, David P, Myers, Vincent, & Silvious, Miranda Schatten. (2009). Mine classification with imbalanced data. *Geoscience and Remote Sensing Letters, IEEE, 6*(3), 528-532.

Williams, Phil, & Norris, Karl. (1987). *Near-infrared technology in the agricultural and food industries*: American Association of Cereal Chemists, Inc.

Wilson, Dennis L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*(3), 408-421.

Wilson, Jeanna L. (2010). Predicting Fertility, Section 1.: Breakout fertility. In M. R. Bakst & J. A. Long (Eds.), *Techniques for semen evaluation, semen storage, and fertility determination* (pp. 72-78): The Midwest Poultry Federation, Buffalo, Minnesota.

Witten, Ian H, Frank, Eibe, Hall, Mark A, & Pal, Christopher J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Wold, S, Johansson, E, & Cocchi, M. (1993). PLS: partial least squares projections to latent structures.

Wold, S, Sjöström, M, & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems, 58*(2), 109-130.
Wold, Svante. (1976). Pattern recognition by means of disjoint principal components models. *Pattern recognition, 8*(3), 127-139.
Wold, Svante, Antti, Henrik, Lindgren, Fredrik, & Öhman, Jerker. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent laboratory systems, 44*(1), 175-185.
Wold, Svante, Esbensen, Kim, & Geladi, Paul. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems, 2*(1-3), 37-52.
Wu, Yingquan, Ianakiev, Krassimir, & Govindaraju, Venu. (2002). Improved k-nearest neighbor classification. *Pattern recognition, 35*(10), 2311-2318.
Xia, Jianguo, Broadhurst, David I, Wilson, Michael, & Wishart, David S. (2013). Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics, 9*(2), 280-299.
Xia, Jianguo, & Wishart, David S. (2011). Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nature protocols, 6*(6), 743.
Xia, Jianguo, & Wishart, David S. (2016). Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis *Current Protocols in Bioinformatics* (pp. 14.10.11-14.10.91): John Wiley & Sons, Inc.
Xie, Jigang, & Qiu, Zhengding. (2007). The effect of imbalanced data sets on LDA: A theoretical and empirical analysis. *Pattern recognition, 40*(2), 557-562.
Xu, Qing-Song, & Liang, Yi-Zeng. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems, 56*(1), 1-11.
Yanenko, Larisa, & Velikanov, Alexander. (2014). Hyperspectral imaging for food quality analysis and control. In N. U. o. F. Technology (Ed.), *National University of Food Technology*. Ukraine.
Yang, Pengyi, Xu, Liang, Zhou, Bing B, Zhang, Zili, & Zomaya, Albert Y. (2009). *A particle swarm based hybrid system for imbalanced medical data sampling.* Paper presented at the BMC genomics.
Yen, Show-Jane, & Lee, Yue-Shi. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset *Intelligent Control and Automation* (pp. 731-740): Springer.
Yen, Show-Jane, & Lee, Yue-Shi. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications, 36*(3), 5718-5727.
Yoon, Kihoon, & Kwek, Stephen. (2005). *An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics.* Paper presented at the Hybrid Intelligent Systems, 2005. HIS'05. Fifth International Conference on.
Yoon, Kihoon, & Kwek, Stephen. (2007). A data reduction approach for resolving the imbalanced data issue in functional genomics. *Neural Computing and Applications, 16*(3), 295-306.
Yoon, Seung Chul, Lawrence, Kurt C, Smith, Douglas P, Park, Bosoon, & Windham, WR. (2008). Bone fragment detection in chicken breast fillets using transmittance image enhancement. *Transactions of the ASAE (American Society of Agricultural Engineers), 51*(1), 331.

Youn, Seongwook, & McLeod, Dennis. (2007). A comparative study for email classification. *Advances and innovations in systems, computing sciences and software engineering*, 387-391.

Yu, Chong Ho. (2000). *An overview of remedial tools for collinearity in SAS.* Paper presented at the Proceedings of 2000 Western Users of SAS Software Conference.

Yu, Miao, Naqvi, Syed Mohsen, Rhuma, Adel, & Chambers, Jonathon. (2011). *Fall detection in a smart room by using a fuzzy one class support vector machine and imperfect training data.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.

Zhang, Minghua, Qin, Zhihao, Liu, Xue, & Ustin, Susan L. (2003). Detection of stress in tomatoes induced by late blight disease in California, USA, using hyperspectral remote sensing. *International Journal of Applied Earth Observation and Geoinformation, 4*(4), 295-310.

Zhang, Wei, Pan, Leiqing, Tu, Kang, Zhang, Qiang, & Liu, Ming. (2014). Comparison of spectral and image morphological analysis for egg early hatching property detection based on hyperspectral imaging. *PloS one, 9*(2), e88659.

Zheng, Zhaohui, Wu, Xiaoyun, & Srihari, Rohini. (2004). Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter, 6*(1), 80-89.

Zhou, Zhi-Hua, & Liu, Xu-Ying. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Knowledge and Data Engineering, IEEE Transactions on, 18*(1), 63-77.