Mortality Prediction based on Imbalanced New Born and Perinatal Period Data

Wafa M. AlShwaish[1], Maali Ibr. Alabdulhafith[2]
College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

***Abstract*—This study was carried out by the New York State Department of Health, between 2012 and 2016. This experiment relates to six supervised machine learning methods: Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF), Deep Learning (DL) and the Ensemble Model, all of which are used in the prediction of infant mortality. This experiment applied ensemble model that concentrated on assigning different weights to different models per output class in order to obtain a better predictive performance for infant mortality. Efforts were made to measure the performance and compare the classifier accuracy of each model. Several criteria, including the area under ROC curve, were considered when comparing the ensemble model (GB, RF and DL) with the other five models (SVM, LR, DL, GB and RF). In terms of these different criteria, the ensemble model outperformed the others in predicting survival rates among infant patients given a balanced data set (the areas under the ROC curve for minor, moderate, major and extreme were 98%, 95%, 92% and 97% respectively, giving a total accuracy of 80.65%). For the imbalanced dataset, (the areas under the ROC curve for minor, moderate, major and extreme were 98%, 98%, 99% and 99% respectively, giving total accuracy increased to 97.44%). The results of the experiments used in this dissertation showed that using the ensemble model provided a better level of prediction for infant mortality than the other five models, based on the relative prediction accuracy for each model for each output class. Therefore, the ensemble model provides and extremely promises classifier in terms of predicting infant mortality.**

Keywords*—Component; machine learning; support vector machine; logistic regression; gradient boosting; random forest; deep learning; ensemble model

I. Introduction

In high-income countries, a significant part of public spending is committed to prevention and care. Chronic illnesses such as cancer, asthma and high fevers present serious barriers to infant survival, and dramatically increase spending on healthcare services. A World Health Organization survey in 2017 estimated that children under five years accounted for 5.4 million of these deaths. The inpatient discharge information has been play a crucial role in improving understanding of risk factors mechanisms in infants. Recent approaches using data mining techniques and machine learning algorithms have become part of the experimental processes of many disciplines over recent years [1].

In the healthcare field, data mining techniques help researchers analyze very large databases, and are useful in directing hospital policies to increase patient flow and minimize non-value-added care time. Classifications based on statistical analysis and Artificial Neural Networks, as well as other Machine Learning algorithms are becoming more common aspects of predictive healthcare models [3]. To increase the accuracy in prediction of machine learning models, new variables relating to patient information are used to construct the models. This study applies machine learning in an effort to change the current healthcare process from a receptive model into one that is increasingly proactive. The research question to be answered in this study can be stated as:

RQ1: Can we identify the features that help to predict infant mortality?

However, no study has yet applied ensemble machine learning methods that assign different weights to different models per output class.

II. Review of Existing Literature

A. Infant Mortality

Identifying the variables which affect statistics on health can be used to predict and thereby address and improve long-term survival rates for infants. Kong et al. (2016) identified various predictors of mortality and morbidity among infants, pinpointing many factors relating to pre-term births aside from gestation and birth weight that could be associated with risks relating to high mortality and morbidity risks. For example, infants born at 24 and 25 weeks were more likely to die than infants born more than 26 weeks into the pregnancy. Identifying infants' risk levels by predicting future health outcomes helps improve the efficiency and quality of health care [2].

Diagnoses relating to the level of infant risk are important in terms of both clinical decision making and the provision of care for newborn children a study by Martinez (2017) identified predictors for prolonged hospitalization or readmission for acute lower respiratory infections (ALRIs) in infants with bronchopulmonary dysplasia (BPD). This wide-ranging study was conducted using nationally representative data from children on a US inpatient database, and included a total of 138 patients. The study used logistic regression with and without an interaction term between gender and breastfeeding. The results of the regression showed a p value of ≤ 0.05 and odds ratios (OR) at 95% confidence intervals [3]. Studies have also proved that lower neonatal mortality rates were associated with early breastfeeding compared with higher mortality rates for late breastfeeding [4].

B. Machine Learning

The following will provide an overview of the various methods that are widely classified as supervised and unsupervised machine learning techniques in predicting infant mortality risks.

Unsupervised machine learning in used to collect data with similar attributes into groups. Sample testing is then classified based on proximity within these groups. The groups are generated based on similarity scales such as probabilistic or Euclidean distance. Ravishankar and Clarke (2017) described commonly used clustering techniques and applied the iterative k-means clustering algorithm, in which the outliers in the legend are used to efficiently identify clusters on Dept. of Health of New York State. The algorithms proved successful in terms of processing for data analysis framework by applying data cleansing/ETL, data joining, classification and prediction, visualization of results, interpretation and reporting. Outliers in cost increases (such as the Monroe Community Hospital) were identified through iterative k-means clustering [5].

Supervised classification techniques are the most commonly implemented methods employed by intelligent systems, and are applicable to cases which contain labeled data. Kenley and Shimony (2016), provide insights into predicting the brain maturity of infants and adolescents using structural and functional magnetic resonance imaging data. Data were evaluated throughout the duration of the functional magnetic resonance imaging of 50 new births from Louis Children's Hospital Neonatal Intensive Care Unit (NICU). The results of the experiment showed that using the Support Vector Machine (SVM) model to achieve results improved accuracy, sensitivity, specificity and p-values for binomial probabilities [6]. Previous studies used data from NICU patient data systems. Rinta-koski and Simo (2017), explored powerful mortality classification models using the SVM and Gaussian Process (GP) to identify clinical features on arrival at the Neonatal Intensive Care Unit and those made during the first 72 hours of care for 598 Very Low Birth-Weight infants (birth-weights of below 1500g), with combined features extracted from sensor measurements. The SVM achieved better classification accuracy (0.931) [7].

III. Experimental Design and Methodology

This section sets out the nature of the experiments that will be used to answer the research question. The CRISP-DM methodology offers a structured approach to data mining [8]. The study will be performed as a five-stage process including evaluation, data understanding, data preparation, modelling and evaluation. Each step in the study will be undertaken using Python programming language and the Tensorflow library [9], an open source library for fast numerical computing, and the Scikit-learn library [10], an open source machine learning library for the Python programming language characterized by several classification, regression and clustering algorithms. The section is divided into sub-sections based on the CRISP-DM framework as shown in Fig. 1, each of which will cover the framework in more detail.

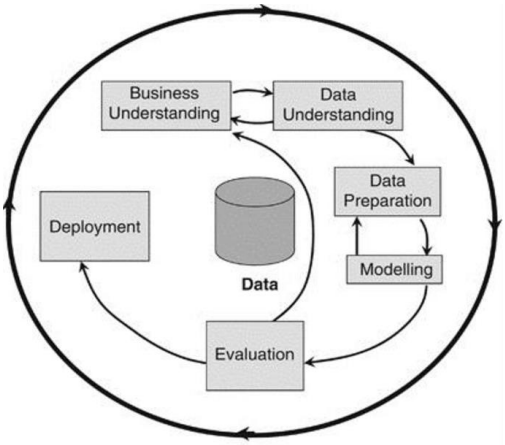

Fig. 1. CRISP-DM Model.

The high-level experiment is illustrated in below Fig. 2.

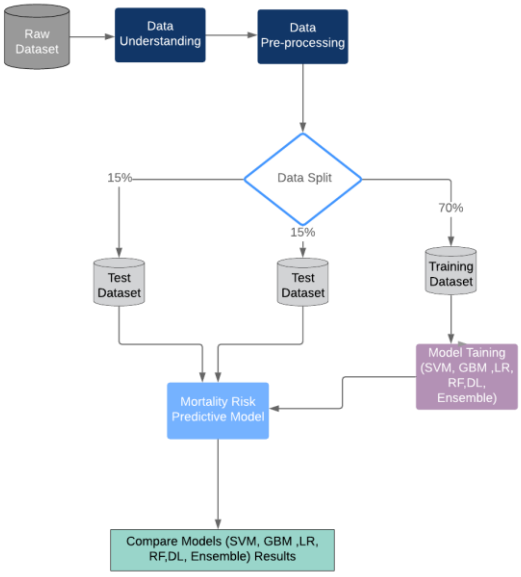

Fig. 2. High Level Design Experiment.

IV. Implementation

This section describes the results of the study and the experiments that were performed. The section pattern emulates that of the Design and Implementation section to make comparisons easier between balanced and unbalanced dataset outcomes.

A. Exploratory Analysis of Dataset

This section explores the infant and perinatal period dataset. First, we need to eliminate the duplicate 'Live born' words from the CCS diagnosis description feature since all patient infants were born alive, and the word does not add much meaning. It can therefore be safely ignored. However, we explored this feature to check the effectiveness of it on each class of mortality. Fig. 3 shows that there are differences between diagnoses in each class of mortality, although the Minor class is similar to entire population because this class represents 97% of data set.

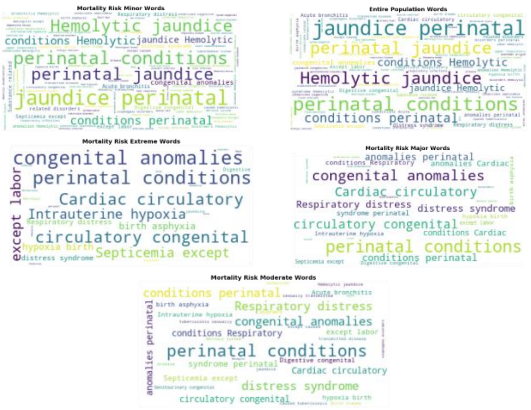

Fig. 3. Entire Population CCS Diagnosis Description Words.

As Fig. 4 shows, the mortality risk distribution by gender indicates that women enjoy better healthcare because the Extreme and Major figures are below the average figure for all patients. Meanwhile, males take less health care because their Extreme and Major cases are above the average patient care data.

However, Fig. 5 and 6 shows that white patients enjoy better healthcare than other races because extreme, moderate and major figures are less than the average in terms of patient care. In addition, the broad range of other ethnicities generally has worse than average health. This appears to indicate discrimination in healthcare depending on race and ethnicity in the USA [11].

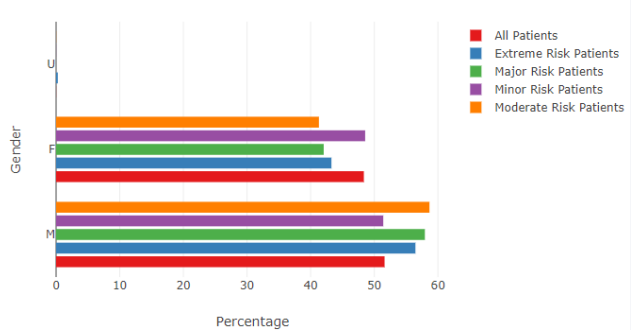

Fig. 4. Mortality Risk Distribution by Gender.

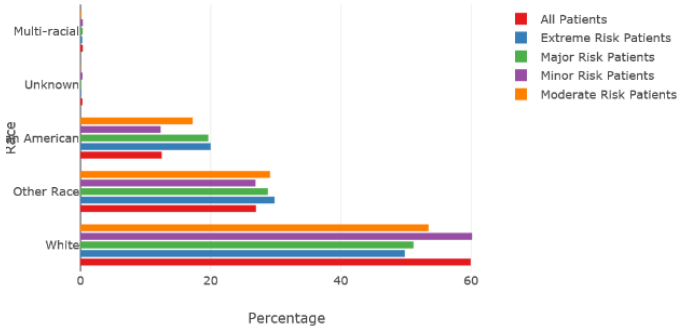

Fig. 5. Mortality Risk Distribution over Race.

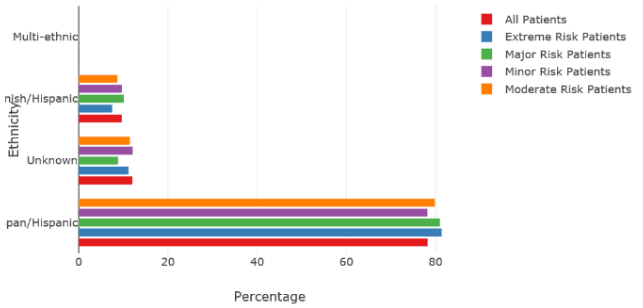

Fig. 6. Mortality Risk Distribution over Ethnicity.

B. Understanding the Data

Statistics were generated to analyse the data in each column to discover a number of different values for each column. From the analysis, it was immediately apparent that the Age Group variable is not important as there is only one value (0 to 17) in the dataset. We will therefore not use this column. In the dataset, the birth weight column has a zero value for some records where we need to analyse mortality rates for new born infants. We checked to see why one row includes zero values for the Birth Weight column. It appears to indicate that a zero birth weight entry means that the patient might not be a new born, but we need to investigate this more deeply to confirm the assumption. After checking the Diagnosis column which also includes records with a zero birth weight value, we observed different diagnoses. That mean records have had zero birth weights entered by mistake, so we need to remove these zero values. However, in order to analyse missing data, we found missing data in only two features (see Table I), so we distinguish those features by assigning a value of -1.

Before we proceed with the analysis, we need to compare different kinds of dimensionality reduction for plotting purposes, since we have 16 dimensions that we need to reduce to 2 dimensions in order to obtain good reduction plotting. Because most of the data is dense, we used principal component analysis (PCA) and t-SNE. Fig. 7 shows the PCA results, in which point distribution is not clear.

However, the t-SNE analysis shown in Fig. 8 shows Extreme cases as clear areas, thereby offering better results than PCA. We then took the process through different kinds of outlier detection algorithms to check outliers from the data such as Robust Covariance, One-class SVM and Isolation Forest. We began with the Robust Outlier detector shown in Fig. 9, in which all points outside the boundary ellipse are outliers.

TABLE. I. MISSING VARIABLE

Variables	Count
Operating provider license number	45681
Other provider license number	159058

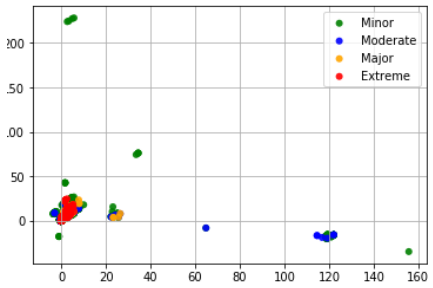

Fig. 7. PCA Data Distribution Plot.

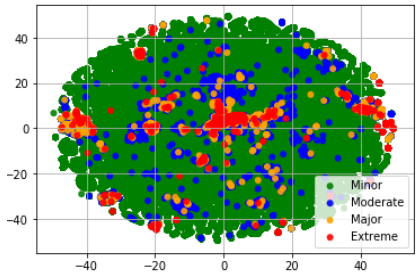

Fig. 8. t-SNE Data Distribution Plot.

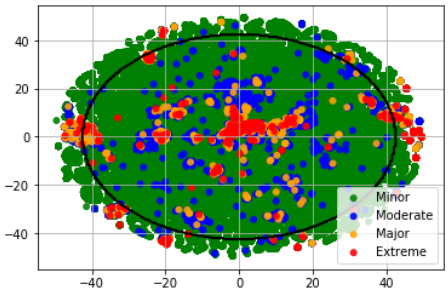

Fig. 9. t-SNE Data Distribution Plot with Robust Outlier Detector.

We then used the Isolation Forest algorithm shown in Fig. 10 which defines a separate boundary between points from outliers of all classes.

However, we needed to perform another kind of outlier detection test. This was the One-class SVM shown in Fig. 11 that sets regional boundaries; all points inside boundaries are valid data, while those outside it is outliers. The One-class SVM was the best detector of outliers, and data within the boundaries are consistent and closed.

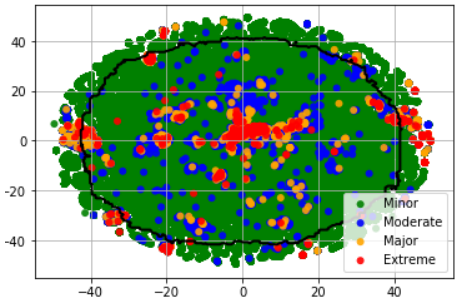

Fig. 10. t-SNE Data Distribution Plot with Isolation Forest Outlier Detector.

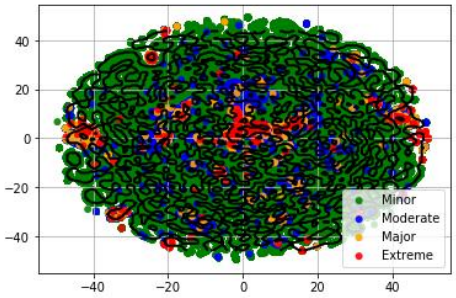

Fig. 11. t-SNE Data Distribution Plot with One-Class SVM Outlier Detector.

Cramer's V Association was then used to interpret associated factors between the nominal variables. The association range lies between 0 and 1, and greater values show stronger associations. The correlation matrix for Cramer's V Association heatmap matrix is shown in Fig. 12, which shows the relationship between the features. The result obtained from matrix is as follows:

- No single feature is strongly associated with APR risk of mortality.
- Patient Disposition and APR Severity of Illness Code have strong positive associations with APR risk of mortality.

The uncertainty coefficient was also used in the study to explain associations between categories. The correlation coefficient determines the degree of association between two variables, and this is shown in Fig. 13, which shows the relationship between features. As shown in both associations the APR Severity of Illness description has a strong positive association with the APR risk of Mortality.

C. Data Preparation

After the analysis of the data is complete, the next step is to remove those issues that have been identified in the dataset so that the remaining data will fit the processes used in modelling.

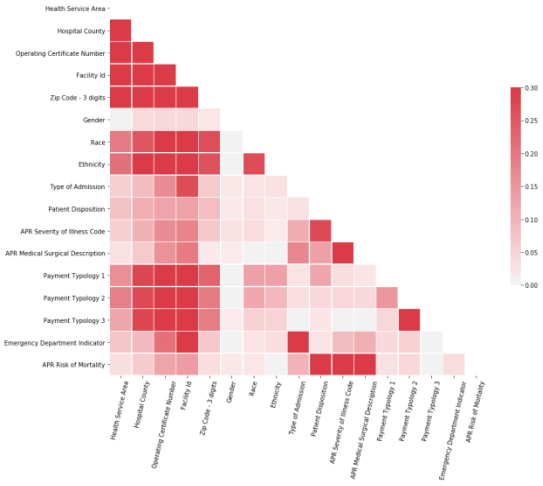

Fig. 12. Cramer's V Association Matrix of Variables with APR risk of Mortality.

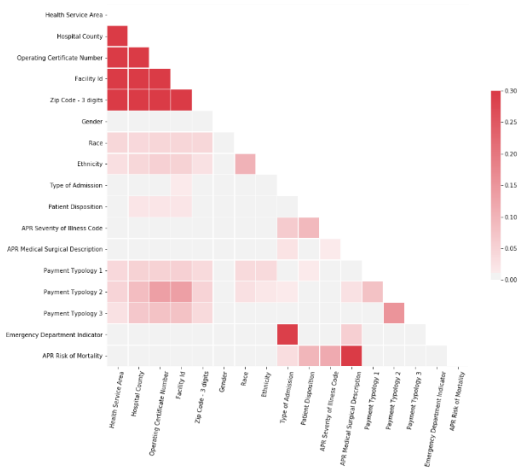

Fig. 13. Uncertainty Coefficient of Variables with APR Risk of Mortality.

1) Balanced dataset: Under Sample:Classifier machine learning algorithms such as Random Forest tend to give results biased towards classes which have the highest number of records.

Classifier algorithms can ignore the features of minimal class, considering them no more than noise. It is highly probable that minimal classes will be misclassified when compared to better populated ones. The reason for using 'Pandas sample' is because we have imbalance between mortality classes, as shown in Fig. 14. The 'Pandas sample' is implemented on the imbalanced dataset samples to balance it. The number of records showing extreme mortality risk is much lower than numbers in the other classes, as shown in Table II.

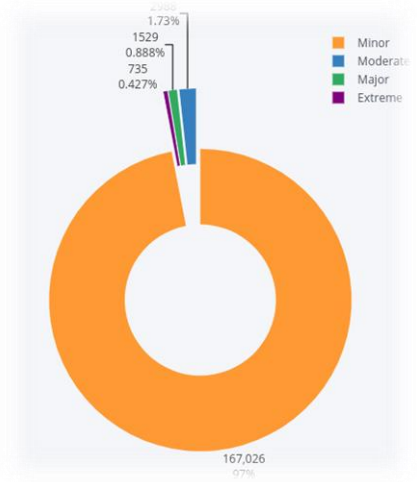

Fig. 14. Unbalanced Data Set.

TABLE. II. UNDER SAMPLING

Target: Mortality Risk	Imbalanced Data set	Balanced Dataset
Minor	167026	735
Moderate	2988	735
Major	1529	735
Extreme	735	735

2) Standardization standard scaler: The distribution of the Birth Weight feature as shown in Fig. 15 provides information about infants' weights. As shown in histogram, the average birth weight is 3.5 kilograms, and birthweights go up to 5 kilograms. Because of this, we need to test for normality using a variety of statistical analyses.

First, we used the Shapiro-Wilk test, which returned a p-value of 0.00, which is less than .05. We then used the Normal-t test, which also returned a value of 0.00. We also used the Anderson-Darling test to see if our data came from a normal distribution. The null hypostudy was rejected, similar to the previous two tests.

The QQ plot could provide us with more certainty about the normality, and also offers better visualization. From the QQ plot shown in Fig. 16 we can see how the data appears, and it is immediately apparent that the data are not normally distributed. This visualization helps us to study abnormal cases in our experiment.

3) Encoding categorical variables: After a balanced dataset had been successfully created, the only problem that remained was to remove categorical variables, as most machine learning models work only on numeric variables and cannot compute using features containing string values. All 18 of the categorical variables were nominal, and this meant that the values within those categorical variables did not follow a specific natural order. In order to remove nominal variables, we performed encoding procedure.

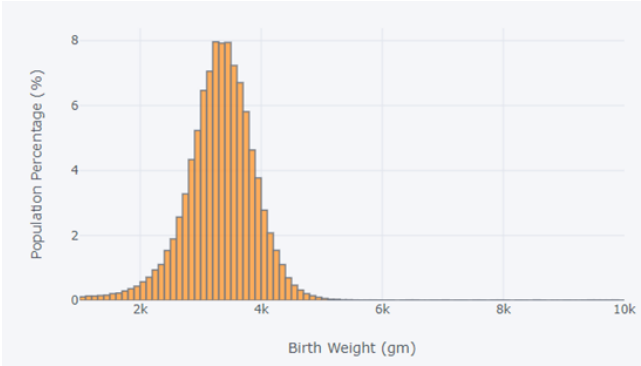

Fig. 15. Entire Population Birth Weight.

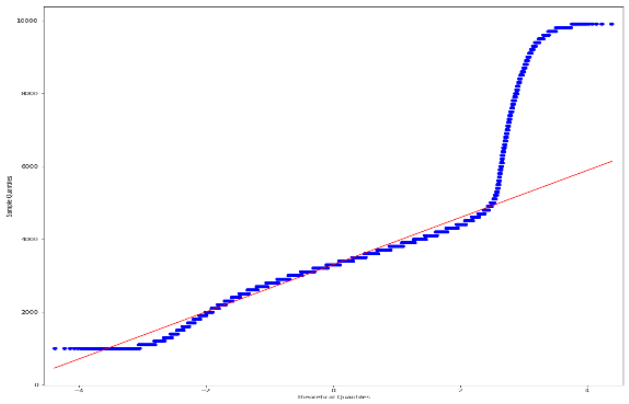

Fig. 16. QQ Plot for Birth Weight.

4) Text vectorization: This part of the analysis focused on creating data vectors from text vectorization by importing a TF-IDF Vectorizer from sklearn.feature_extraction.text. The vectorizer was initialized, fitted and transformed to calculate the TF-IDF score for the text in the [(CCS Diagnosis Description)] feature. The sklearn fit_transform performed both fit and transform functions, and the output took the form of a skewed matrix.

D. Models

In this phase, we built classification models to predict infant mortality risk using Gradient Booted (GB), Support Vector Modelling (SVM), Random Forest (RF), Logistic Regression (LR), Deep Learning (DL) and ensemble models. The balanced data set was created after adding and encoding categorical data and normalizing the results to use for the construction of models. Before training the model, input data regarding text vectorization and strength of association of features was implemented to obtain a better fit for the model. For each model we processed data in two ways: first, we used the scikit-learn train_test_split method to divide the dataset into training, validation and testing datasets that would maintain the distribution of the output. 70% of the data were used for training, 15% for validation and 15% for testing.

1) Logistic regression: Logistic Regression was used to provide a multi-class classification regression model. First, the LR module was imported to create an LR classifier object using the Logistic Regression cross validation function to get best parameters.

2) Gradient boosted tree: We used XGBoost to predict the mortality risk for infants. We imported XGBoost, which uses an assessment metric to check the performance of the Training Model on the test dataset.

3) Random forest: The Random Forest (RF) model used a randomized search function to evaluate the best hyper-parameters. While the parameters were learned during the model training, hyper parameters must be set before training. The importance of each feature in the RF classification is indicated by the sum of the reduction in Gini Impurity (a measure that the decision tree uses to minimize when splitting each node for every node that is split by that feature. We can use these to attempt to calculate which of the predictor variables the RF considers most important in terms of mortality risk. The feature importance can be extracted from a trained RF.

4) Support vector classifier model: We applied a prepackaged model provided by a scikit-learn support vector classifier to train an SVM model on this data. Tuning parameter values for machine learning algorithms effectively improves the performance of the model.

5) Deep learning: We used fully connected network architecture to implement our infant mortality risk prediction model. We used the Class Weight variable to calculate the class weight and added it to model. After the model had been created, we were able to make predictions according to all the learned nodes.

6) Ensemble model: Most of the known ensemble techniques do not account for the relative prediction accuracy of multiclass classification problems. It either uses a blanket weighting for all classes or uses voting, which could lead to equal votes for multiple different outputs, as shown in Fig. 17.

In our approach, we decided to feed into the deep neural network the output probability per class from the different models, as they would allow the deep neural network to give different weights to different models per output class. This improved the overall prediction accuracy, which was based on the relative prediction accuracy for each model per output class as shown in Fig. 18.

One limitation in most previous studies is that they only considered a blanket weighting for all classes, and no previous study has concentrated on assigning different weights to different models per output class.

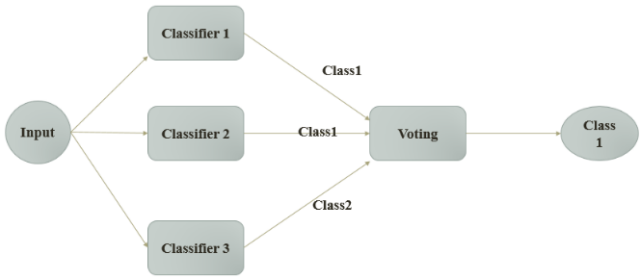

Fig. 17. Voting Mechanism.

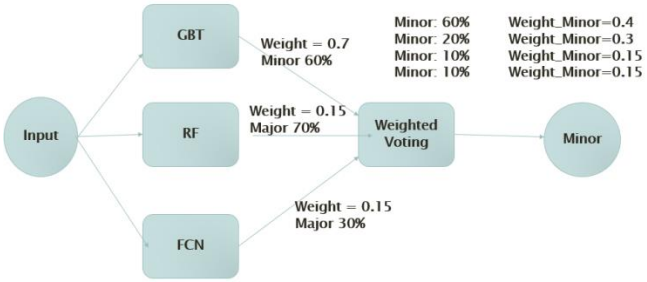

Fig. 18. Voting Mechanism with different Weights.

V. Evaluation and Results

A detailed analysis of the experiments described in the previous section will be provided in this section, including the results of each experiment. The experiments were performed in order to build six models for supervised machine learning. This section evaluated the execution of each model according to the levels of accuracy gained after running each experiment on the dataset. The same experiment was also performed on the imbalanced dataset which contained biased values relating to the mortality risk features. When evaluating the performance of models, box plots were created using confusion matrixes that summarized the prediction results generated as well as the accuracies achieved. We also used cross validation techniques by applying a series of training/validation/test set splits based on logistic regression and random forest methods. The statistical analysis of the study result will also be discussed in this section.

A. Comparison of Average Performance of Imbalanced and Balanced Target Data

Imbalanced data sets are common in predictive classification experiment. Fig. 19 shows the results for each classifiers on the imbalanced dataset in which the mortality risk was highly biased towards instances having a 'minor' classification, for all models.

The first analysis was performed on the results gained from the unbalanced data set. It is clear from the histogram of the accuracies that random forest and the ensemble model have higher accuracies than the models derived from the other algorithms. The maximum F1 score obtained by random forest model for the 'extreme' class is above 60%.

Due to the imbalanced data, the under-sampling approach has been taken to create a balanced dataset. Using alternative metrics like F1 score, recall, precision, true positive rate and false positive rate is strongly recommended in place of using the accuracy of the model to measure its performance.

Tables III and IV shows the mean accuracies and classification metrics obtained from the imbalanced dataset and balanced data set of each model.

B. Comparison of Classifiers Performance

A further experiment was performed on all six models after applying the under-sampling technique. Fig. 20 shows the performance histogram of balanced dataset models. The most remarkable change here is the increase in F1 score for each target variable value. As the graph shows, the random forest model and the ensemble model have higher prediction accuracy when compared to other models.

C. Strengths and Limitations of Results

Machine learning algorithms and their use was considered as an integral factor in the research. The experiment used six machine learning algorithms (LR, RF, GB, DL, Ensemble and SVM), which were similar in the ways in which they were used for the classification of variables.

The training models that were relevant to the different families in the same data set can be seen as one of the strengths of the study. Likewise, the ensemble model was built using different models–random forest, gradient boosting and deep learning–in order to obtain an efficient performance from the model. In the ensemble model, we decided to feed the output probability per class from the different models into the deep neural network, as they would allow the network to assign different weights to different models per output class. This improved the overall prediction accuracy, which was based on the relative prediction accuracy for each model per output class. The experiment gave us the opportunity to compare six models, which meant that the results obtained are more important than the results obtained by comparing only two models.

The experiment also concentrated on analysing the impact of balancing a data set that was initially unbalanced. We used under-sampling to remove bias from the results, and a significant improvement on the performance of all models was achieved by applying the under-sampling process. Techniques relating to data pre-processing–such as feature scaling using z-scores and converting categorical to numeric variables–were studied in detail during the experiment and subsequently applied to the data in order to improve the outcomes.

As far as the limitations of the experiment are concerned, the study was based on records relating to patients from a particular hospital, and may therefore have been biased towards the population of a specific region. Additionally, the time span used for monitoring the patients was small (5 years). To provide improved forecast results, the period of observation should be increased in order to obtain comparatively stable values, and the result of this might have an effect on predictive modelling results.

D. Summary of Analysis

The results and evaluation of the research has been discussed in this section. All six models were built on two data sets, one with biased values in terms of the target variables and one with balanced values. Cross-validation was performed using LR and RF to obtain optimal parameters in order to enhance the models' performance. The ensemble models (RF, GB and DL) outperformed the RF, SVM, LR, GB and DL models in the prediction of mortality risk for both the balanced data set (Total Accuracy 80.65%) and the imbalanced data set (Total Accuracy 97.44%). In the ensemble model, we applied a new approach that no study has previously attempted by feeding the output probability per class from each model into the deep neural network, as this network would assign different weights to different models per output class. This improved the overall prediction accuracy in our experiment, which in turn was based on the relative prediction accuracy for each model per output class. We can therefore recommend this approach in other areas that have multiclass classification problems. The result also indicated a weaker performance of the DL model on a balanced dataset (Total Accuracy 70.89%) than on an imbalanced dataset (Total Accuracy 83.56%).

The strengths and limitations of the results concentrate on the data pre-processing techniques, which were used to improve models performance. The concluding section which follows will offer a detailed summary of the study, as well as participation and effects, and will also offer avenues for further research.

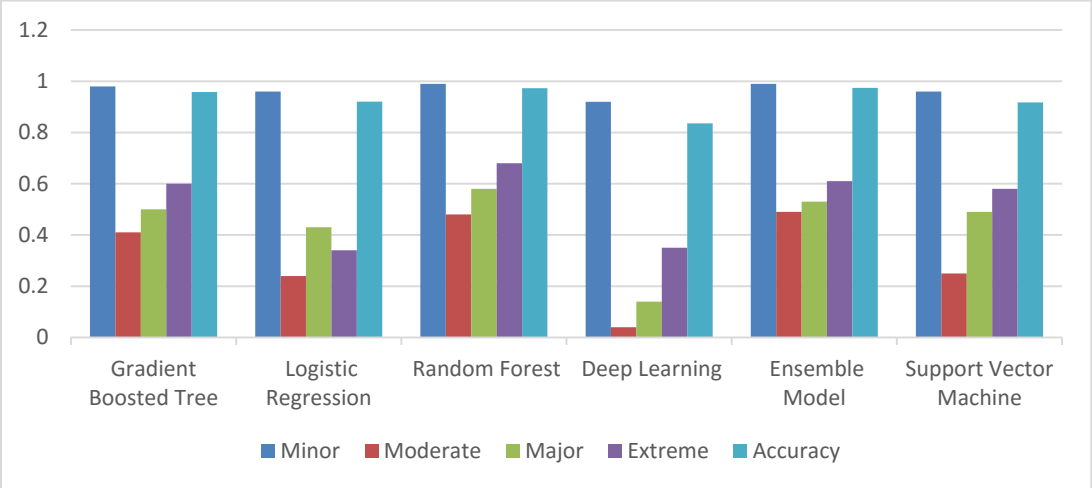

Fig. 19. Model Comparison: Imbalanced Dataset.

TABLE. III. PERFORMANCE IMBALANCED TARGET (MORTALITY RISK)

	Accuracy	Precision				Recall				F1 Score			
		Minor	Moderate	Major	Extreme	Minor	Moderate	Major	Extreme	Minor	Moderate	Major	Extreme
Gradient Boosted Tree	95.76%	1.00	0.28	0.49	0.56	0.97	0.78	0.50	0.65	0.98	0.41	0.50	0.60
Logistic Regression	92.03 %	1.00	0.15	0.37	0.22	0.93	0.58	0.52	0.71	0.96	0.24	0.43	0.34
Random Forest	97.26%	1.00	0.39	0.55	0.72	0.98	0.63	0.61	0.59	0.99	0.48	0.58	0.68
Deep Learning	83.56%	1.00	0.03	0.07	0.31	0.85	0.10	0.93	0.40	0.92	0.04	0.14	0.35
Ensemble Model	97.44%	0.99	0.40	0.59	0.55	0.99	0.63	0.48	0.70	0.99	0.49	0.53	0.61
Support Vector Machine	91.70%	1.00	0.15	0.46	0.54	0.93	0.75	0.53	0.64	0.96	0.25	0.49	0.58

TABLE. IV. PERFORMANCE BALANCED TARGET (MORTALITY RISK)

	Accuracy %	Precision				Recall				F1 Score			
		Minor	Moderate	Major	Extreme	Minor	Moderate	Major	Extreme	Minor	Moderate	Major	Extreme
Gradient Boosted Tree	76.34	1.00	0.52	0.63	0.78	0.91	0.70	0.63	0.73	0.95	0.60	0.63	0.75
Logistic Regression	71.87	0.99	0.52	0.56	0.71	0.90	0.85	0.48	0.63	0.94	0.64	0.52	0.67
Random Forest	79.09	1.00	0.67	0.65	0.77	0.92	0.67	0.67	0.82	0.96	0.67	0.66	0.80
Deep Learning	70.89	0.95	0.50	0.52	0.71	0.94	0.67	0.46	0.66	0.94	0.57	0.49	0.68
Ensemble Model	80.65	0.99	0.62	0.69	0.76	0.91	0.73	0.63	0.81	0.95	0.67	0.66	0.78
Support Vector Machine	70.99	0.93	0.44	0.57	0.86	0.86	0.67	0.67	0.58	0.89	0.53	0.62	0.69

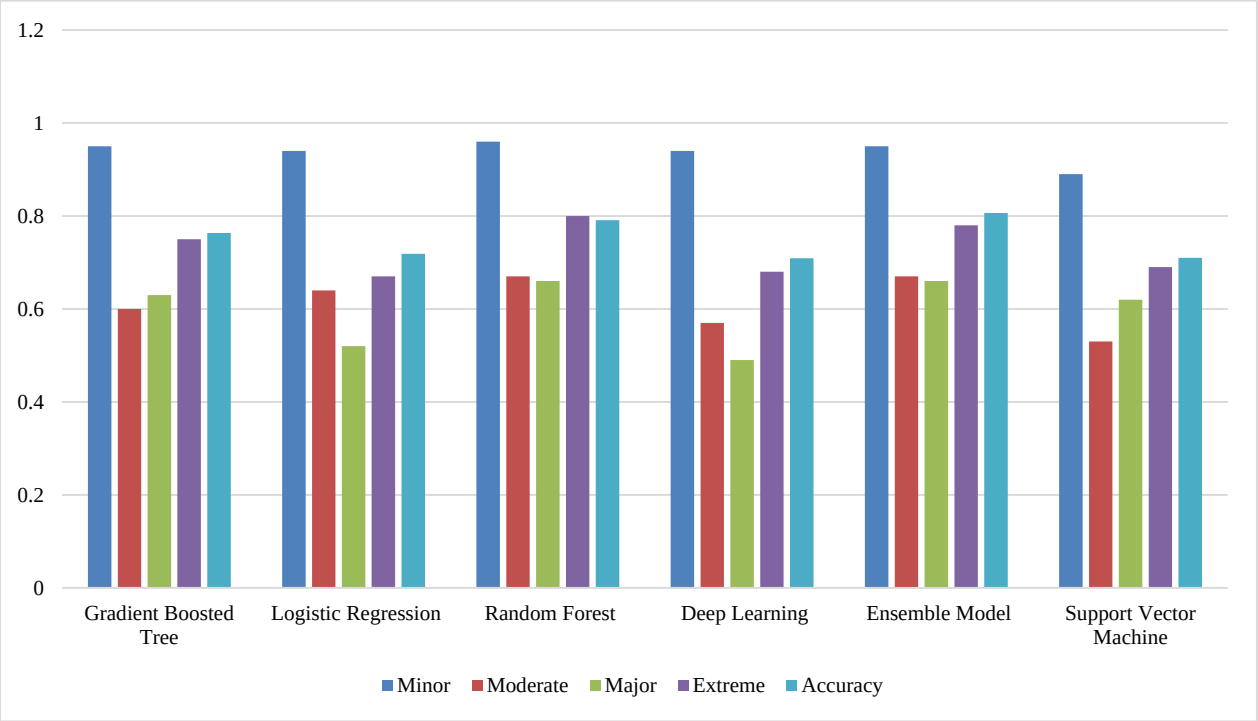

Fig. 20. Model Comparison: Balanced Dataset.

VI. Conclusion

A. Research Overview

This dissertation took the form of an investigation of multiple supervised machine learning techniques. These techniques were used to analyse factors relating to discharge details concerning infant patients. The work offered a literature review that summarised existing studies into both machine learning and mortality prediction. An experiment using six supervised classification techniques was performed in order to construct predictive mortality risk models using data that had been collected from the New York State Department of Health's state wide planning and research cooperative system over a five-year period. The first used the Support Vector Machine (SVM) technique to classify supervised learning techniques, creating a model using line or hyperplane data. Logistic Regression (LR) is a more traditional predictive technique in medical study, where the probability of multi classed occurrences is examined. The Gradient Boosting (GB) is a powerful technique for building predictive models that include weak learners, loss function and the additive model. Random Forest (RF) is a prediction algorithm which is used to run test feature data through randomly created trees. At the first connected layer, each neuron in the learning algorithm receives input from all factors from the previous layer. Finally, the Ensemble method combined several machine learning techniques (in this case DL, RF and GB) into a single predictive model in order to improve prediction levels All techniques could be used to predict mortality risks for infants using patients' information in different ways. The main purpose of the study was to measuring the model accuracy and the F1 score, and to compare the performance of each model in order to conclude which model offers the best performance in terms of prediction accuracy.

B. Problem Definition

The limitations identified in the existing literature and gaps in the research were used as motivation for the dissertation. Rinta-koski and Simo (2017) suggest that more promising methods such as SVM can be used to identify clinical features on arrival at the Neonatal Intensive Care Unit, as well as features observed during the first 72 hours of care for 598 Very Low Birth-Weight infants. However, Ahmadi et al (2017) applied Random Forest techniques to survey maternal risk factors that were associated with low birthweight neonates, using data mining on information collected from Milad Hospital to account for interactions between variables. The most commonly used algorithm to identify diseases is logical regression, so comparisons of accuracy were made between Logistic Regression, Support Vector Machine, Random Forest, Deep Learning, Gradient Boosted Tree and the Ensemble model.

The experiment was performed to empirically determine which of the six classifiers offers the better performance, giving a positive answer to the research question asked at the start of the dissertation, which was "Can we identify which features help to predict infant mortality?" No study has yet applied ensemble machine learning methods concentrating on assigning different weights to different models per output class in order to obtain a better predictive performance for infant mortality.

C. Future Work and Recommendations

This project focused only on patients from a particular hospital, and might have biased towards the population of a

specific region. Further research should be conducted on monitoring and capturing more patient information from hospitals in different regions or countries, which would help build a more generalizable model. There were important variables that could not be considered in this experiment, including the mother's age, which could be useful for analyzing a new approach to create labels using three categories (18 to 28, 29 to 39 and 40 to 49) in an attempt to identify relation between the age of the mother and infant mortality risk. These data should also be collected and analysed to increase prediction accuracy.

REFERENCES

[1] J. Xu, S. L. Murphy, K. D. Kochanek, E. Arias, and D. Ph, "Mortality in the United States , 2015," no. 267, pp. 1–8, 2016.

[2] X. Kong et al., "Neonatal mortality and morbidity among infants between 24 to 31 complete weeks : a multicenter survey in China from 2013 to," BMC Pediatrics, pp. 1–8, 2016.

[3] C. E. Rodriguez-Martinez, R. Acuña-Cordero, and M. P. Sossa-Briceño, "Predictors of prolonged length of hospital stay or readmissions for acute viral lower respiratory tract infections among infants with a history of bronchopulmonary dysplasia," Journal of Medical Virology, vol. 90, no. 3, pp. 405–411, Mar. 2018.

[4] L. C. Mullany et al., "Breast-Feeding Patterns , Time to Initiation , and Mortality Risk among Newborns," no. April, pp. 599–603, 2018.

[5] A. R. Rao and D. Clarke, "An open-source framework for the interactive exploration of Big Data : applications in understanding health care," pp. 1641–1648, 2017.

[6] C. D. Smyser et al., "NeuroImage Prediction of brain maturity in infants using machine-learning algorithms," NeuroImage, vol. 136, pp. 1–9, 2016.

[7] O. Rinta-koski and S. Simo, "Gaussian process classification for prediction of in-hospital mortality among preterm infants."

[8] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME : Data Mining Methodology for Engineering Applications-A Holistic Extension to the CRISP-DM Model ScienceDirect DMME : Data Mining Methodology for Engineering Applications – A Holistic Extension to the CRISP-DM Model," 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, vol. 96, no. July, 2018.

[9] M. Abadi et al., "TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning," 2016.

[10] G. Moncecchi, "Learning scikit-learn : Machine Learning in Python."

[11] R. A. Hummer, "Black-white differences in health and mortality: A review and conceptual model," Sociological Quarterly, vol. 37, no. 1, pp. 105–125, 1996.