REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

USING DEEP LEARNING METHODS TO CLASSIFY LITHOLOGY WITH MULTI-LABEL IMBALANCE DATA

Eman IBRAHIM ALYASIN

Doctor of Philosophy

Supervisor

Assoc. Prof. Dr. Oğuz ATA

İstanbul, 2024

USING DEEP LEARNING METHODS TO CLASSIFY LITHOLOGY WITH MULTI-LABEL IMBALANCE DATA

Eman IBRAHIM ALYASIN

Electrical and Computer Engineering

Doctor of Philosophy

ALTINBAŞ UNIVERSITY

2024

The dissertation titled USING DEEP LEARNING METHODS TO CLASSIFY LITHOLOGY WITH MULTI-LABEL IMBALANCE DATA prepared by EMAN IBRAHIM ALYASIN and submitted on 13/06/2024 has been **accepted unanimously** for the degree of PhD in Electrical and Computer Engineering.

Assoc. Prof. Dr. Oğuz ATA

Supervisor

Thesis Defense Committee Members:

Assoc. Prof. Dr. Oğuz ATA	Department of Software Engineering, Altınbaş University	_____
Assoc. Prof. Dr. Sefer KURNAZ	Department of Computer Engineering, Altınbaş University	_____
Assoc. Prof. Dr. Doğu Çağdaş ATİLLA	Department of Electrical and Computer Engineering, Altınbaş University	_____
Prof. Dr. Hasan Hüseyin BALIK	Department of Computer Engineering, İstanbul Aydın University	_____
Assoc. Prof. Dr. Aytuğ BOYACI	Department of Computer Engineering, National Defense University	_____

I hereby declare that this dissertation meets all format and submission requirements for a PhD dissertation.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Eman IBRAHIM ALYASIN

Signature

ABSTRACT

USING DEEP LEARNING METHODS TO CLASSIFY LITHOLOGY WITH MULTI-LABEL IMBALANCE DATA

IBRAHIM ALYASIN, Eman

Ph.D., Electrical and Computer Engineering, Altınbaş University

Supervisor: Assoc. Prof. Dr. Oğuz ATA

Date: June / 2024

Pages: 59

To anticipate and detect lithologies in a variety of surveys, geologists can save operational expenses and uptime by utilising deep learning methodologies and applications. Accurate data processing and scientific research using data gathered in different geological areas are made possible by this. The four lithologies data in the present research were analysed and classified using multi-class imbalance issues and high dimensionality. One of the biggest issues facing modern data analysis is the imbalance in data classification. Particularly when combined with other challenging factors like the existence of overlapping class distributions, and data imbalance can have a significant impact on the accuracy of classification. When there are several classes involved, mutual imbalance relationships between them exacerbate the situation, making its influence more evident. Furthermore, the high dimensionality issue may result in overfitting and increased computational complexity, both of which may impair classification efficiency. Recursive Feature Elimination (RFE) is used to find the most valuable predictive features, while Synthetic Minority Oversampling (SMOTE) is used to resample the data. Using hybrid multi-class DL system unbalanced learning approach is our solution to solving these issues. Finally, by offering precise categorization and quick responses about the interpretation of data collected in many study regions, we think that our innovations might contribute to the advancement of geological research.

Keywords: Lithology, Hyperparameters, Feature selection, Optimization, SMOTE.

ÖZET

ÇOK ETİKELİ DENGESİZLİK VERİLERİ İLE LİTOLOJİYİ SINIFLANDIRMAK İÇİN DERİN ÖĞRENME YÖNTEMLERİNİN KULLANILMASI

IBRAHIM ALYASIN, Eman

Doktora, Elektrik ve Bilgisayar Mühendisliği, Altınbaş Üniversitesi

Danışman: Doç. Dr. Oğuz ATA

Tarih: Haziran / 2024

Sayfa: 59

Çeşitli araştırmalarda litolojileri tahmin etmek ve tespit etmek için jeologlar, derin öğrenme metodolojileri ve uygulamalarını kullanarak operasyonel masraflardan ve çalışma süresinden tasarruf edebilirler. Doğru veri işleme ve farklı jeolojik alanlardan toplanan veriler kullanılarak bilimsel araştırma yapılması bu sayede mümkün olmaktadır. Mevcut araştırmadaki dört litoloji verisi, çok sınıflı dengesizlik sorunları ve yüksek boyutluluk kullanılarak analiz edilmiş ve sınıflandırılmıştır. Modern veri analizinin karşılaştığı en büyük sorunlardan biri veri sınıflandırmasındaki dengesizliktir. Özellikle örtüşen sınıf dağılımlarının varlığı ve veri dengesizliği gibi diğer zorlayıcı faktörlerle birleştirildiğinde, sınıflandırmanın doğruluğu üzerinde önemli bir etkiye sahip olabilir. Birkaç sınıf söz konusu olduğunda, aralarındaki karşılıklı dengesizlik ilişkileri durumu daha da kötüleştirerek etkisini daha belirgin hale getirir. Ayrıca, yüksek boyutluluk sorunu aşırı uyum ve artan hesaplama karmaşıklığıyla sonuçlanabilir ve bunların her ikisi de sınıflandırma verimliliğini olumsuz etkileyebilir. Özyinelemeli Özellik Eliminasyonu (RFE), en değerli tahmine dayalı özellikleri bulmak için kullanılırken Sentetik Azınlık Aşırı Örnekleme (SMOTE), verileri yeniden örneklemek için kullanılır. Hibrit çok sınıflı DL sistemi dengesiz öğrenme yaklaşımını kullanmak bu sorunları çözme çözümümüzdür. Son olarak, birçok çalışma bölgesinde toplanan verilerin yorumlanmasına ilişkin kesin sınıflandırma ve hızlı yanıtlar

sunarak, yeniliklerimizin jeolojik araştırmaların ilerlemesine katkıda bulunabileceğini düşünüyoruz.

Anahtar Kelimeler: litoloji, Hiperparametreler, Özellik seçimi, optimizasyonu, SMOTE.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
ANN	:	Artificial Neural Network
CNN	:	Convolutional Neural Network
DL	:	Deep Learning
DLMs	:	Deep Learning Models
ML	:	Machine Learning
RNN	:	Recurrent Neural Network
SMOTE	:	Synthetic Minority Oversampling Technique

1. INTRODUCTION

1.1 BACKGROUND

The most common technique for teaching computers new information is Machine Learning (ML). For the past ten years, computers were mostly used for training purposes and data storage. By combining the study of computers with the quick analysis of massive amounts of data, it is now feasible to improve algorithms and their processing exponential due to the quick advancement of computing Scientists and the public disclosure of new technology [1].

ML finds applications in speech recognition, Natural Language Processing (NLP), population estimation, health condition detection, porosity computation in sedimentary rock recognition of minerals in sedimentary stones. The method of ML gives answers from extensive data analysis and predicts decision outcomes with high accuracy [2].

The controlled learning approach constructs a model for predicting new information based on the supplied reference data. ML algorithms require data in standard formats and kinds, and data with accurate and precise values acquired from reliable resources, to improve the prediction process. Several of the major accomplishments of computer science now employ Deep Learning (DL) in alongside ML as an emerging methodology. Two essential elements of Multi-Layered ML algorithms are pattern recognition and Mineral Prospective Mapping (MPM) [3], [4], [5].

Multi-layered network learning has made DL algorithms dominant when working with large data sets for classification and forecasting. Convolutional Neural Network (CNN) is a type of Feed-Forward Neural network, for example, convolutional processing and Deep-Structure. Because of their combined weight design and translation stability, CNNs are useful in mineral exploration as they can identify the inner connections of different geological features, discover hidden mineral information, and discriminate models that would otherwise be missed by traditional ML approaches.

Accurate data categorization improves work, improves search results, and automates jobs in less time and with fewer employees. Precision computational techniques and tools assist geologists in identifying lithology in offshore well surveys by allowing better data analysis. ML and geology are integrated with scientific technological research, research formats, and

technique to allow model recognition in sedimentary rocks. If a physical sample of sedimentary stones is required, the expert will need visual explanatory materials as well as laboratory studies. This is a lengthy and costly procedure that requires hiring highly skilled and dedicated staff. Moreover, the physical features of sedimentary stones can be used to detect rocks instantly in the real environment, saving time and money. This approach has been used to calculate litho phages over seismic intervals [6], to map geological and mineral surfaces[7], and to quantify permeability using drill profiles [8].

In recent years, the classification of lithologies has become a popular study topic, and various studies have used ML and Data Mining (DM) approaches to solve the problem [9], [10], [11]. Indeed, there are two basic ways to identifying lithologies in the literature: Supervised, and Unsupervised approaches. Supervised ML methods based on measurements of physical characteristic are used to predict different types of rocks. It uses physical property measurements and scale type information to classify works, as evidenced by the term "Supervised". Random Forests [12], XGBoost [13], and AdaBoost[14] are three well-known techniques for controlled rock classification. Algorithms built for the unsupervised approach aim to detect hidden patterns in unmarked data. We identify the SOM approach (self-organized maps) and the K-means problem clustering method as examples of the unsupervised methods used [15], [16], [17].

Regardless of the efforts made to categorise rocks, there are still many obstacles in this subject, which makes work in it extremely difficult. For example, the lithologies classification dataset is severely unbalanced. When an individual rock type's data set is significantly greater than those of other rock groups, the algorithms are incorrectly classified. However, traditional measures of assessing rock models, such specificity and accuracy, are only adequate for determining the validity of predictions; they do not provide enough information to determine how well the model operates in terms of lithological class classification. Where, a measure of accuracy, focuses in the prediction process on the class that has a high data distribution (majority class) [18]. F1-score [19], [20] is a more realistic indicator of measuring the predictive efficacy of a petrology test by combining sensitivity with measures of accuracy. The F1-score evaluates the classification test prediction and balances the usage of sensitivity and accuracy.

Several research projects in literature have employed models without using feature selection techniques. Results could be erroneous if all characteristics are classified accurately [21]. Therefore, increasing rocks classification sensitivity or decreasing misclassification may not come from employing all features without first identifying the highly impacted features for prediction [22], [23]. In summary, choosing predictive characteristics is essential to creating a successful rocks detection model. By integrating the resilience of several ML approaches and methodologies, we present a hybrid model in this study that combines both supervised and unsupervised learning to increase rock prediction efficiency. The Grid-Search optimisation method was utilised to estimate the hyperparameters of our model, and a Recursive Feature Elimination (RFE) was employed to choose the predicted attributes.

1.2 PROBLEM FORMULATION

As shown in Table 3.1 the dataset groups have 5 different classes of Lithology with uneven distribution of samples. Whereas the GP dataset has a large number distribution with class (10, 14, 15) in contrast to the class number (11, 12, 13) since it has a small number of sampling distribution. This confirms that the distribution of Lithology classes dataset is highly imbalanced. Common ML methods provide excellent prediction accuracy over the majority class but poor prediction accuracy over the minority class when learning from imbalanced data. In another hand, traditional models evaluate model performance with a measure of accuracy, and this will lead to inaccurate classification of unbalanced data sets, where accuracy is highly dependent on prediction with classes of data that have a large distribution of data. Numerous DL models have been developed to address high-dimensional multi-class unbalanced classification to enhance overall performance.

1.3 OBJECTIVES

The following is a summary of the primary study goals:

In this work, lithological data from offshore wells was classified using ML and DL methods to create a new system.

It has also been demonstrated that DL techniques outperform traditional ML techniques in the classification of lithology. It has also been demonstrated that hybrid DL strategies outperform conventional DL techniques in lithology classification systems.

Moreover, the efficiency of the lithological classification model was improved by employing the SMOTE oversampling strategy for data balance.

To minimise dataset feature irregularities and expedite processing, Wrapper Feature Selection (WFS) has been applied. Consequently, every model's output showed a general increase in performance. Instead of utilising X-ray images, provide predictive research with multiple DL models utilising numeric datasets of IODP offshore drilling log characteristics.

1.4 THESIS OUTLINE

The following sections are a description of the dissertation details:

Chapter 2: Outlines the most recent studies as well as background descriptions of the Lithology data sets, feature selection and imbalanced dataset.

Chapter 3: Presents the proposed methodology design and implementation, as well as the recommended methodologies for SMOTE and the feature selection model for multiclass imbalance data.

Chapter 4: The experimental results of the proposed method were presented, beginning with DL model experiments, and ending with proposed method experimental results.

Chapter 5: Finally, we discuss our findings and recommendations to the future work.

2. LITERATURE REVIEW

2.1 INTRODUCTION

An essential component of geological research is lithologic prediction, which calls for the examination of sedimentary, reservoir, and stratigraphic facies [24]. Reservoir data is gathered, real-time drilling monitoring is done, and reservoir descriptions are made using lithology [25]. Moreover, one factor influencing the time and techniques of subsurface support for engineering is lithology [26], [27]. Accurate and timely rock stone identification may decrease construction hazards, avert tunnel collapse, and enhance technical feasibility during tunnel excavation and well drilling. As a result, strategies for quickly and accurately identifying rock stones must be studied. In recent years, several research have used ML and DL technologies to assist advance the discipline of geology by providing real - time responses.

The author [28] used ML methods to a supervised stone classification problem on a broad scale, employing remote sensing and limited geophysical data. Random forests techniques are commonly employed for multi-class inference utilizing publicly accessible, multi-source, high-dimensional geophysical data because they are easy to handle, robust across a large range of models hyperparameters, and efficient. They are also used when encountering spatially dispersed training data. The author concludes that the random forest approach is a great tool for creating accurate first predictions for real geological mapping applications.

In another study, the authors [29] gave a number of ML methods for oil discovery and engineering, for classifying subterranean shale formations from drilling data, and for ML based on Tree-based ensemble models. The authors attempted to improve the predictability of their ensemble model by applying several ML ensemble models (AdaBoost and Logit Boost meta-algorithms) to two data sets from Daniudui Gas Field (DGF) and Hangjinqi Gas Field (HGF). Several metrics for calculating accuracy, memory, and F1-scores were evaluated by dividing the data into 5-fold, and 10-fold cross-validation, and the Logit Boost algorithm showed the highest performance.

In this work the author [30] employed two datasets in this study: Conventional Wireline Logging (CWL) data and Logging While Drilling (LWD) data from the Yan ‘an Gas Field. The main goal was to compare three ML algorithms: Random Forest (RF), One-Versus-One

Support Vector Machines (OVO SVMs), and One-Versus-Rest Support Vector Machines (OVR SVMs), to improve a more practical way in the field for LWD systems. To reduce the dimensionality of the input data, a correlation analysis was performed on the registration data. The ideal hyperparameter for each algorithm and 10-fold validation technique was determined using grid search optimization. The overall evolution performance metrics with RF shown the high accuracy with 90%. As a result, the OVR SVM and RF classifiers result in fewer prediction errors than the OVO SVM classifiers in the lithology classification evaluation matrix.

Maxwell et al. [31] Lithotypes were dynamically classified from geophysical log data using ML algorithms. The author used geological log data and machine learning (ML) approaches, such as gradient boosting algorithms and random forests, to predict modified and non-altered lithotypes because altering coal projections have gotten little study. The Bowen Basin of Eastern Australia provided the author with 1230 samples from 263 boreholes in a huge data collection. Eighty percent of the data from the training and test splits is used to predict the alteration class, and the remaining twenty percent is utilised to evaluate each ML model to select the model that performs the best.

Another study [32] employed pixel-based image analysis to enhance land cover classifications in Iran. Multi-sensor layers gather a large amount of data. The lithology result is more accurate than the information gathered. Several ML models were used to evaluate performance. RF and SVM were selected as the best classifications for data analysis of images, with accuracy rates of 98% and 97%, respectively.

Bressan et al. [33], The International Oceanographic Program(IODP) multivariate logarithm data from offshore wells was processed using a ML technique. The IODP trip was separated into four templates to enhance the outcomes of the practical approach and the study of the range of expeditions. The four ML models were trained, validated, and tested using cross-validation. RF and MLP were shown to be the most efficient methods for finding every template.

2.2 FEATURE SELECTION

Reliability selection and multi-class imbalanced learning have been extensively studied in the last few decades. Whereas [34] thoroughly analyses unbalanced learning strategies

including data sampling and cost-sensitive learning methods, [35] covers feature selection techniques like filtering, wrapping, and embedding methods. An overview of research on selecting features and unbalanced learning is provided in this section.

Because high-dimensional features are sometimes prone to overfitting, computing complexities, and prediction problems, feature selection methods were developed to eliminate irrelevant and redundant features, therefore increasing classification accuracy, and minimizing computation time. A substantial number of researchers [36], [37] have demonstrated their efficiency and efficacy by analysing high-dimensional data and improving learning outcomes. Feature selection approaches are usually classified into three types based on the interaction between features and supervised learning [38][39] (I) Filter techniques, (II) Wrapper methods, and (III) Embedding methods are all available.

Filter methods, which are distinct of the training process, select subsets of features by comparing sample data's statistical behaviour to predetermined criteria. There are two stages in a typical filtering method. Using feature evaluation criteria, the significance of features is first ranked. Second, just the most important features are chosen, with the rest discarded. Other feature evaluations for filter techniques have been presented, such as feature correlation [39], [40], information gain, maximum information coefficient [41], and others. Filtering algorithms are computationally cheap since they have less direct contact with the process of learning, but also usually result in lower precision in the prediction model because of feature set picked isn't ideal for a particular learning task [40].

Wrapper methods combine a learning algorithm with a selection method to investigate the feature subset combining area and evaluate the quality of candidate subgroups [40]. To put it another way, inside wrapper methods, three components must be described: the classifier, the assessment criteria, and the search methodology. Unlike filter techniques, wrapper approaches encompass training set into the feature set, resulting in the optimal extracted features for the used classifier [42]. Wrapper techniques use more computing resources than filter methods [43] since each subset must be evaluated through a learning process. Wrapper feature selection strategies are commonly employed in feature selection challenges due to their ability to improve classification performance while decreasing the cost of features.

Embedded approaches, in contrast to filtering and wrapping methods, where feature selection is used within the classifier, which is strongly related to the selected learning algorithms. The embedded methodology is also often considered a subset of the wrapper method, which is characterised by a more relationship between feature selection and classifiers construction [40]. An integrated approach has been developed in recent years. Some classifiers, such as: Random Forest (RF), Decision Tree (DT), Classification and Regression Tree (CART), can select functions built into them. In the process of optimizing feature selection for other classifiers that do not have innate abilities, various forms of penalties have been created. Although the Vector Support Machine (SVM) cannot choose the features on its own, numerous SVM-based techniques have been created utilizing various definitions of the penalty term, such as Penalty L1 and Penalty for Fine Absolute Deviation [44]. As can be seen, the built-in technique is class-specific and cannot be extended to other classifiers, severely restricting the choice of classifiers and the analysis of classification issue efficacy.

2.3 IMBALANCED LEARNING

Most of the ML algorithms have greatly influenced their performance in recent years due to unbalanced data sets. In an imbalanced data set, one or more classes will have much more samples than other classes, with majority class having significantly so many samples and the minority class having significantly lesser. Unbalanced learning aims to develop a model that is more successful than the traditional class imbalance classifier in achieving classification accuracy[45]. Several methods have been introduced to deal with class imbalance problems. Data acquisition and amplification are two widely used strategies [46]. Before the training phase, data sampling is used to equalize the uneven collected data by adding or deleting samples from the training set.

Random Oversampling (ROS) is the most basic way of over-sampling, in which samples from the minority-class are randomly recreated, while Random Subsample (RUS) is the easiest form of under sampling method, where samples from the majority-class are randomly deleted. Because of their simplicity, ROS and RUS are routinely used to treat class imbalances. However, ROS reproduces the same examples over and over again, while RUS ignores potential information from most classes, leading to data overload or information loss

[47]. Several exemplary approaches are presented, including heuristic techniques, to reduce ROS and RUS deficiencies.

3. IMPLEMENTATION AND DESIGN MODELS

3.1 METHODOLOGY

The practice of preparing raw data to increase the usefulness and effectiveness of prediction models is known as data pre-processing. We need to solve two issues to get accurate classification with high dimensions imbalanced histological data sets.

Features must first be chosen to determine the best combination of features with the most knowledge and fewer functions. Second, data is resampled using SMOTE to achieve high accuracy in both majorities and minorities. To choose the optimal parameters for deep learning models (DLMs), Grid Search is also employed as Hyperparameter Optimisation (HPO) technique. Because controlling the overall behaviour of the algorithms depends on selecting the appropriate hyperparameter tuning technique. Figure 3.1 shows the suggested model for multiclass unbalanced data.

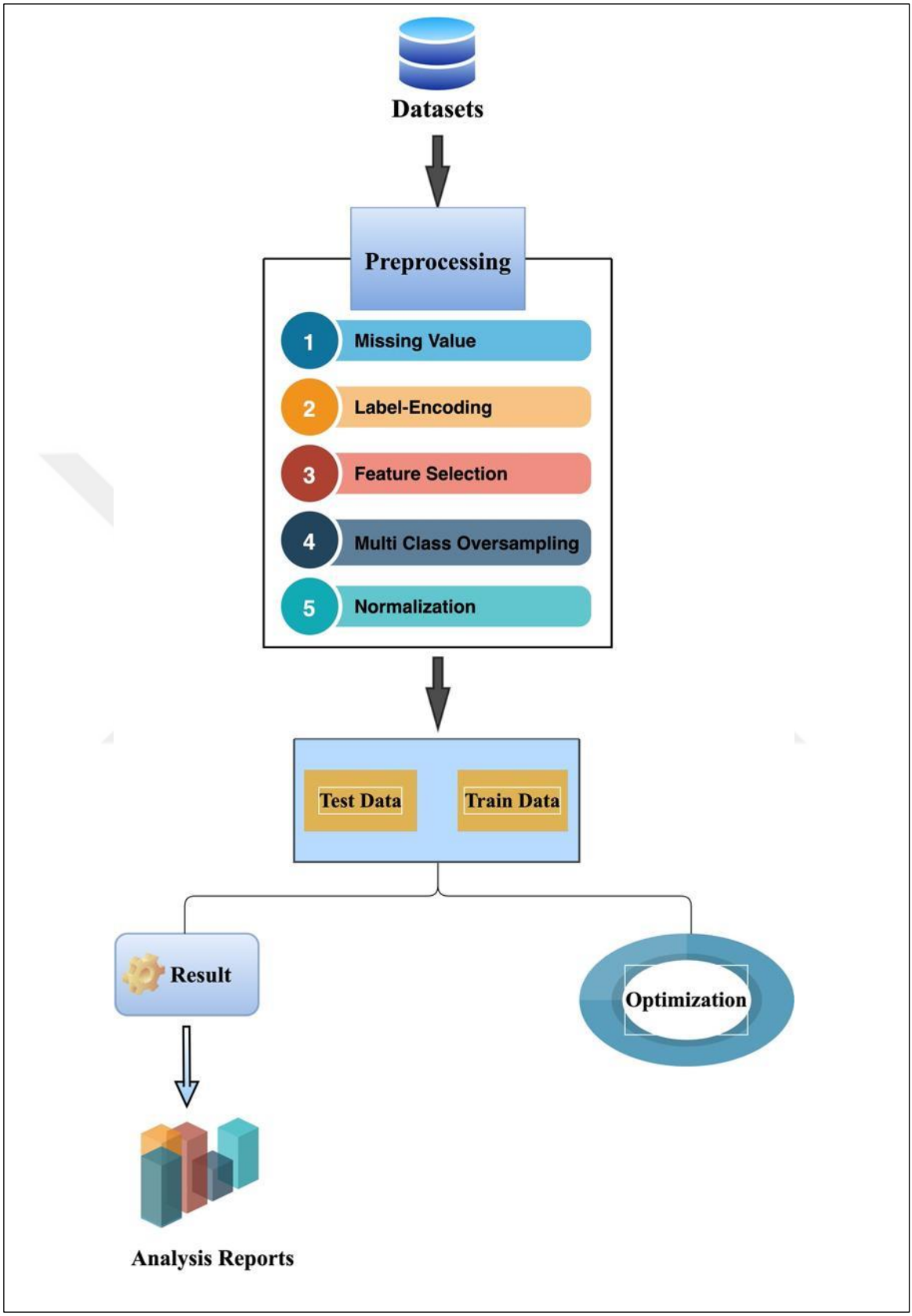

Figure 3.1: Machine Learning Hybrid Model.

3.2 DATASET DESCRIPTION

A few instances of multi-class imbalanced information are provided in this section using four lithology data sets. In addition, the use of three multi-class imbalanced convolution datasets found in the UCI data database demonstrates that the proposed model is not solely influenced by lithology data.

3.2.1 Lithology Datasets

This study's dataset is derived from the IODP's lithology identification method, which uses univariate data for logarithm variables derived from drilling offshore [14]. According to IODP expedition's suggested lithological study. Expeditions (349, 354, 355, 356, 359, 361 and 362) of the IODP were used to create this data collection. In several parts of the Indian Ocean, holes were drilled. During the IODP expedition, the onboard data collected on board the ship provides the first step in working with sample rock. This data is collected in a basic or collective random sample.

The data comprises various classes the lithologies, Table 3.1 illustrates. For instance, the GP dataset comprises 5 classes, class 14 of which includes 9116 samples, and class 11 of which includes just 361 samples overall. The data is deemed multiclass-imbalanced, due to the disparate allocation of the information among the classes.

Table 3.1: Distribution of Data in A Multi-Class Lithology Datasets.

Groups	**Lithology Code**					
	10	**11**	**12**	**13**	**14**	**15**
GP	4778	**361**	1113	176	8960	9116
GP1	4778	-	-	-	10073	9116
GP2	4778	-	-	537	10073	9116
GP3	4802	537	1895	1514	6640	9116

3.2.2 Convolutional Datasets

This study uses three multiclass unbalanced UCI data sets alongside to the lithology data to examine the efficacy of a proposed hyper DL model for handling multidimensional imbalanced learning problems. As Table 3.2 illustrates, these datasets are distinguished by four main attributes: (I) the number of occurrences (IN), (II) the numbers of features (NF), (III) the numbers of classes (NC), and (IV) the class distribution. The datasets span from 1728 examples for automotive evaluation to 7200 examples for thyroid disease, and a total of 358 dermatological cases are covered.

Table 3.2: Convolutional Datasets Description.

Dataset	IN	NF	NC	class distribution
Car Evaluation	1728	6	4	65/69/384/1210
Thyroid-Disease	7200	21	3	66/368/6666
Dermatology	358	33	6	111/60/71/48/48/20

3.3 DATA SCALING

Most machine learning and DL algorithms do not work well with varied numerical data features. Therefore, there are two solutions that are usually used for this problem: Standard-scaler and Min-Max scaler using the python scikit Learn [48] library. The min-max scale defines a range (usually 0-1) and scales in or out to scale digital properties to a predefined range. Furthermore, the standard scaler turns the numerical features into a distributions with the a means of 0 and a standard deviation of 1 [49]. Regular test results x is computed by subtracting the mean from each test score and dividing by the standard deviation, as shown below:

$$Z = \frac{(x - u)}{s} \tag{3.1}$$

where u is the mean of the samples, and s is the standard deviation.

3.4 EVALUATION METRICS

In this study, several performance evaluation measures were used to evaluate the results of all developed deep learning models [33]. The evaluation metrics uses a confusion matrix to predict the correctly and incorrectly evaluation results by using different classification metrics: True-Positive (TP), True-Negative (TN), False-Positive(FP) and False-Negative(FN).

Accuracy: Percentage of all correctly classified samples divided by all correctly and incorrectly classified samples, to determine the expected correct percentage of the classifier, and is represented by Eq.2:

$$Accuracy = \frac{TP+TN}{TP + TN + FP+ FN} * 100 \quad (3.2)$$

Precision: defined as proportion of all properly identified samples (TP) divided by the total number of correctly and wrongly classified samples (TP, FP) , and is shown by Eq.3:

$$Precision = \frac{TP}{TP + FP} \quad (3.3)$$

Recall: It is a percentage of classified samples (TP) divided on the total number of positive classified samples. An essential indicator of the quantity of accurate data anticipated and the impact of information imbalance on outcomes. The recall is computed using Equation 4.:

$$Recall = \frac{TP}{TP+ FN} \quad (3.4)$$

F1-score: It is a key indicator measure for evaluating the effectiveness and efficiency of a model, which combines precision and recall. It is also one of the important metrics that are sensitive to data imbalance and affect the efficiency result is represented by Eq.5:

$$F1\text{-}score = \frac{2 .TP}{2 . TP + FP+ FN} \quad (3.5)$$

For classification problems with unequal class distributions, we aim for a very comprehensive evaluation that takes into account various aspects, including measuring the classifier ability to achieve balance between multiclass-imbalanced datasets [33].Our experiment uses the area under the curve (AUC), F1 score, and the geometric mean (G-mean) because it shows resistance to an imbalanced distribution.

Geometric mean (G-mean): indicates, as shown in Eq. 6, the Square-Root of the value of the combined value of the class specificity and sensitivity in binary identification. This metric aims to maximize each class accuracy while maintaining its balance [47].

$$G{-}mean = \sqrt{Sensitivity\ x\ Specificity} \tag{3.6}$$

The receiver operating characteristic (ROC) is a graphical method for evaluating the classification performance of models. It uses a probability curve containing two parameters, true positive rate (TPR) represented by Eq.7 and false positive rate (FPR) represented by Eq.8, to determine the optimal model to predict the classes.

$$TPR = \frac{TP}{TP+FN} \tag{3.7}$$

$$FPR = \frac{FP}{FP+TN} \tag{3.8}$$

ROC curves plot (TPR versus FPR) at various classification thresholds. Because of its multidimensional capabilities, the result variable may be shown more clearly across the graph spectrum [33]. The classification model, which serves the same role in both classes, is represented by the descending diagonal (0,1).

Further, ROC curve has been extensively used by researchers for its ease of visualization and interpretation[50]. It is also recognized as a trustworthy method for assessing the performance of several ML algorithms. Because we have a multi-imbalance class problem, our research focuses mostly on AUROC values for model comparisons.

3.5 SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

A common problem in machine learning is unbalanced data sets. Binary unbalanced information has two classes, while multiclass imbalanced data comprises more than two classes[47]. As table 3.1 shows, the data we gather has multiple classes imbalanced dataset issues because of a discrepancy in the distribution of the overall amount of class samples—some groups have a significantly lesser frequent than others and they are referred as being minority-classes. However, there is bias in the accuracy of class classification when certain classes—referred to as majority classes—have a significantly higher distribution than others. SMOTE was employed to address the multiclass imbalanced problem that was previously described. Oversampling algorithms increase categorization by raising minority samples

rather than processing majority samples, as opposed to under sampling techniques. In order to produce a new minority sample that performs better in classification, SMOTE [45] fine-tunes a randomly selected, homogeneous surrounding sample. To further increase the number of minority group samples and maintain the class's balance, SMOTE is a potent ML technique whose main idea is to randomly generate new samples from minority group samples and their neighbours [51]. SOMTE is frequently employed in the class imbalance region in various domains, such as network intrusion detection and sentence boundaries in speech, as a result of its significant improvement in the overfitting condition over the past few years owing to the non-exploratory sample selection method [52].

3.6 MULTI-CLASS RECURSIVE FEATURE ELIMINATION

The feature selection procedure is essential for preventing oversampling and improving classification performance when handling a classification problem with many learning samples and multiple dimensions, such as the GP, GP1, GP2, and GP3 datasets. However, it's also crucial to determine which traits are necessary for use to determine which features are most crucial to the categorization problem. As a result, a ML model can function more quickly and effectively if fewer characteristics are used during training [34], [39]. In practice, feature reduction is challenging and usually requires extensive and intricate testing.

Numerous methods for selecting anticipated variables or eliminating non-predictive ones are found in ML research [37], [38]. Among these is a wrapper method called Recursive Feature Elimination (RFE), which is also known as the greedy searching method. These techniques explore every feature combination that can be combined in order to find the feature that performs the best for a ML algorithm. These procedures have the advantage that will ensure that the best set of features is chosen and if done together with cross-validation in general is very powerful for overfitting.

In this work, we employed four datasets each of which has a unimportant number of features. Using the complete data set to train models would be a time-consuming procedure. Therefore, RFE was used as crucial preprocessing step to minimized amount of features in order acquire most significant features as a first stage in the model's learning phase. As a result, the findings are more easily understood, and classification models are more efficient overall.

3.7 DEEP LEARNING ALGORITHMS

In Figure 3.1 illustrates the two phases of DLMs: learning and prediction. In the classification algorithm training step, the properties of the items in every class are established. allowing this relationship model to be created on the classifier's foundation. Models are then used to forecast fresh data during the testing phase. The features of the expected data classes are used to create these models. The following models are utilized to create unbalanced multi-class lithology models in this study:

3.7.1 Artificial Neural Networks (ANN)

ANN is a computer programs that simulate how the human brain and biological cells process abilities and strategies for interpreting information. In the ANN type, artificial neurons can be organized into layers and linked together using connections or nodes. The bias voltage transferred between neurons, the weights associated with neural connections, and the transition function (activation function) that turns the right-hand input into a single output when a bias value occurs are all part of the ANN design. By changing node values that control the direction of knowledge, neurons can be taught to do a specific task [53].

The contribution of each hidden neuron is determined by weighing the neurons of input-layer and relating them to the opposing neurons of hidden-layer, this process is called the dense layer. To maximize network capability for each data collection, number (size) of these hidden-layers and their linked nodes must be determined. Because the related nodes' values are generally different, positive nodes are more important than negative nodes.

As a result, when one node is connected to another, the computer uses the inverse weight to determine which node is the most important [54]. ANN was used to address a range of issues by using the computer to create mathematical simulations to copy nature from the brain. The computer will find the answer to every dilemma just like the human brain by utilizing this algorithm [55]. The mechanism work of ANN show in Figure 3.2.

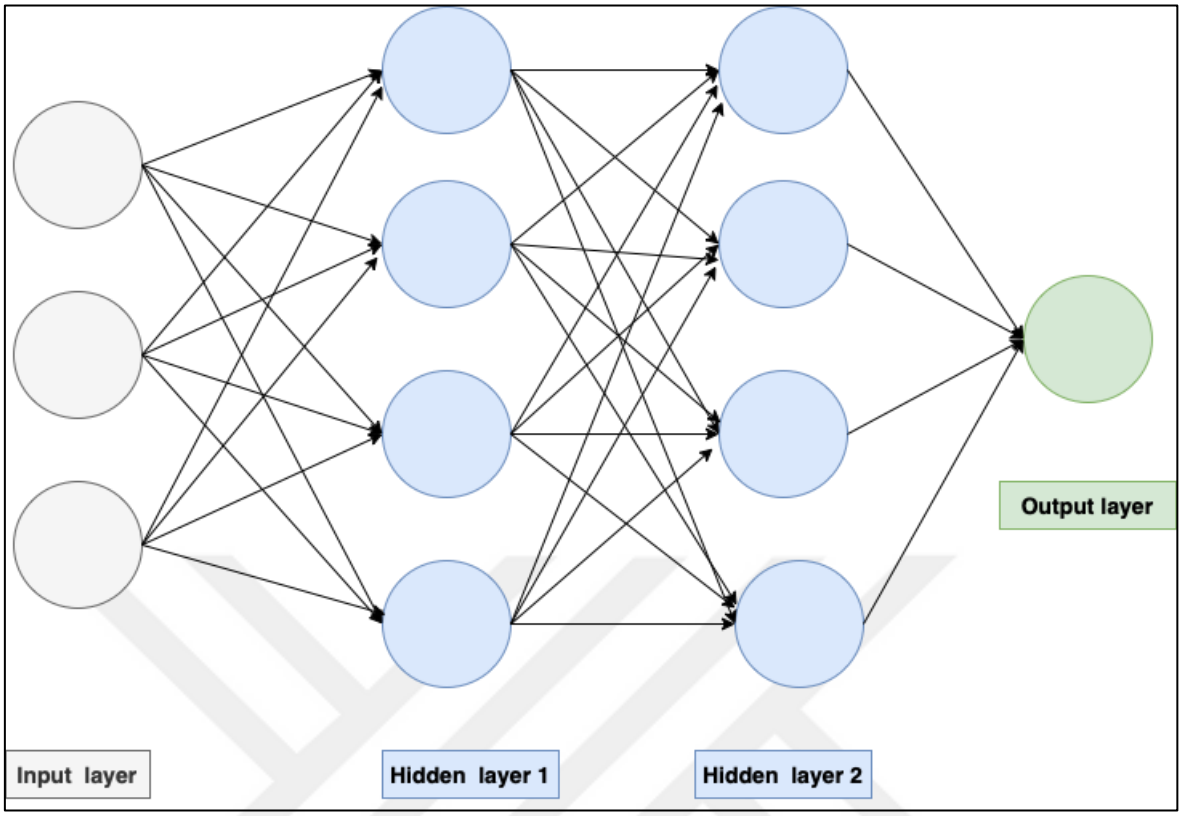

Figure 3.2: ANN Architecture.

3.7.2 Convolutional Neural Networks (CNN)

CNN is unlike any other type –of Neural network. A convolution network is used to accomplish mathematical convolution. It's simply a linear process in which the input data are multiplied by a training data set. The convolutional neural network is built on this concept. This method was created with two-dimensional data principles and is still used for that. It can handle both 1D and 3D data [56]. The MLP approach is a regularized form of the CNN. Each neuron in one layer of an MLP network is linked to all neurons in the layer above it. As a result, there's a better chance that data will be delivered to an MLP network. To fix the situation, CNN took a variety of regulatory procedures. Convolution, activation, pooling, and the fully connected layer are the four layers that make up a CNN. Convolution and grouping layers are mostly used to extract features. The convolution layer conducts linear operations such as convolutions, which contain a series of arithmetic operations. As a result of this layer, feature maps are generated [57] the lower sample layer is also known as the Pθθling layer. The Pθθling layer reduces the amount of feature maps, lowering computational costs and reducing overfitting. Nonlinear functions like Sigmoid and tanh are

employed in the activation Layer. Because it employs the Rectified Linear units (ReLU)function. The outcome is zero if input data are negative numbers, and the function becomes linear if the input values are positive numbers. At the end of each convolutional layer, this procedure is utilized to improve the efficiency of the computational activity[58]. Finally, the data classification process occurs in a fully connected layer. The deviation value is added to this layer after the input value is multiplied by the weight matrix. When the layer is exited, the chances for inputs correctly classified are determined. In addition, before shifting the feature map to a fully linked Layer, the Flattening step converts it into a 1D dimensional array. The mechanism work CNN display in Figure 3.3.

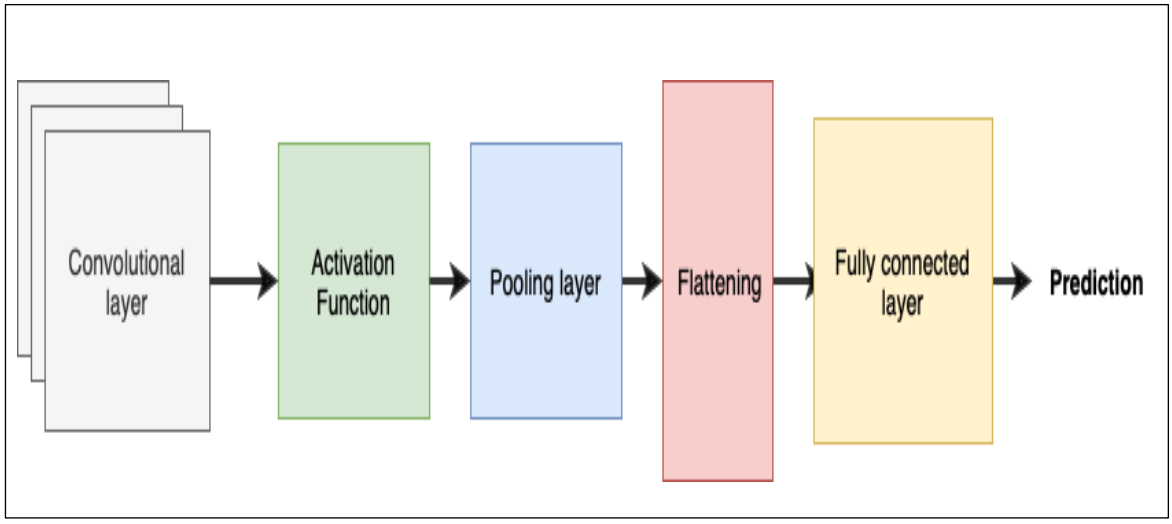

Figure 3.3: CNN Architecture.

3.7.3 Recurrent Neural Networks (RNN)

The brain is essentially a functioning system that heavily relies on repetition in its processes of feeding oneself of actions, perception, education, free and automatic behaviours, and basic learning in living things. Processing power is decided by the redundancy design of networks, which analogous to brain that it has at least a single feedback link. In terms of dynamic system aggregate accuracy, a network that is redundant is substantially more compact than a forward one. The basic function of a RNN is to evaluate-data flow in via a hidden layer unit. The results of the previous series of measurements determine it. The Cohen-Grossberg model and the Hopfield model are two RNN networks groups. Cohen-Grossberg neural networks are a type of neural network that was first proposed by Cohen and Grossberg in 1983[59]. Cohen-Grossberg neurons are a subset of other neural network types, including

cellular neural networks, which include genetic control systems, Hopfield neuronal networks, and bidirectional memory-based neural network [60]. This all demonstrates that Cohen-Grossberg neural systems can produce amazing results. It's worth noting that dynamical properties play a big role in these outstanding achievements (e.g., synchronization, passivity and stabilization) [61]. RNNs rely on sequencing data in numerous fields, including speech recognition, text processing, and the health business, including DNA sequencing and other healthcare applications. The mechanism work RNN display in Figure 3.4.

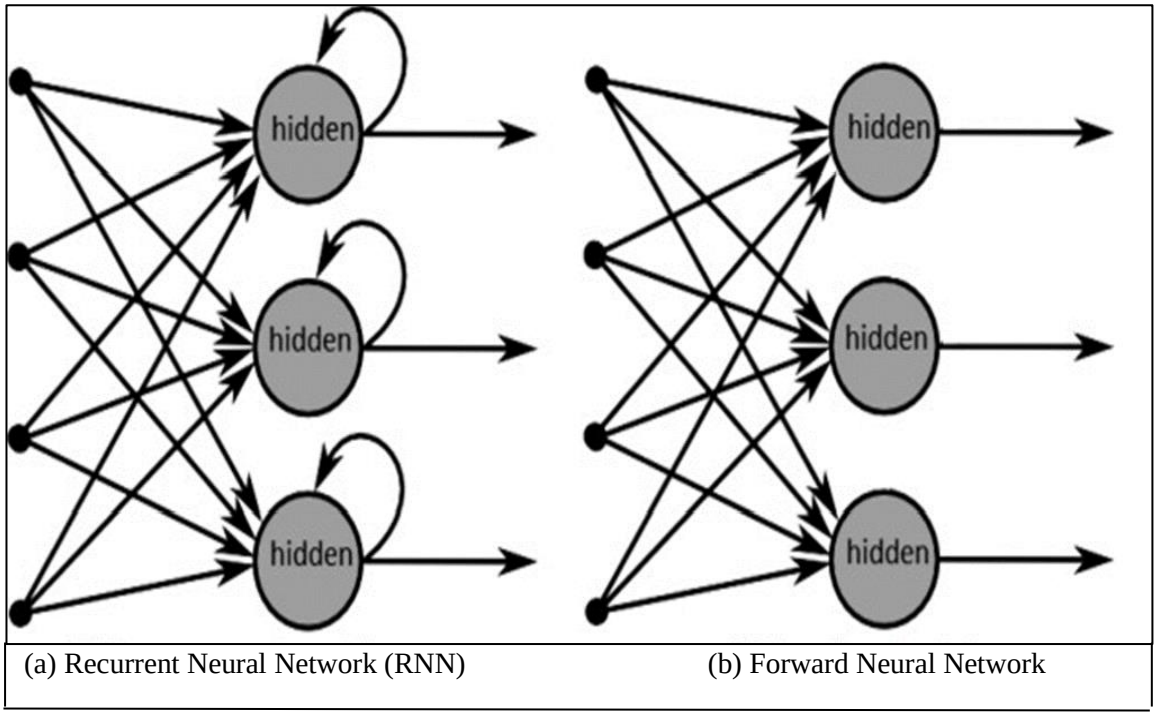

Figure 3.4: RNN Architecture [62].

3.8 HYPERPARAMETER OPTIMIZATION

Selecting the best hyperparameter is a big challenge and calls for trying to find an exchange-of among model generalization and model complexity. The approach of constructing an ideal model architecture for the optimal hyperparameter configuration is known as hyperparameter setup. Hyperparameter optimization is viewed as a vital component in developing an effective DL model[63], particularly for deep neural networks with numerous hyperparameters. Different forms of hyperparameters, such as categorical, discrete, and continuous hyperparameters, trigger different hyperparameter settings in DL algorithms [64].

Manual testing is an old way of modifying hyperparameters that is still used in postgraduate studies, but it necessitates a detailed understanding of the ML algorithm and its hyperparametric value setting. Due to the large number of hyperparameters, dynamic models, time-consuming model calculations, and nonlinear hyperparameter interactions, manual tuning is inefficient for several problems. Before discovering the ideal parameter values, grid-search entails adjusting the value based on the range of hyperparameters and incrementing each hyperparameters with a validated time interval. To discover the best-performing hyperparameters, we did a grid search through a grid of selected hyperparameters in this study. Additionally, optimization was used to decrease prediction cost, error rate, and over-processing, as well as to identify the highest performing Set of parameters.

3.9 PROPOSED METHODOLOGY

This paper proposes a methodology for identifying lithology. The suggested system's main goal is to improve classification metrics by analysing lithology data. The suggested system combines the hybrid CNN-RNN deep learning algorithm with data imbalance processing and feature selection for this goal (RFE). As illustrated in Fig. 3.1, the proposed system comprises of data pre-processing, data balancing, feature selection, and classification sections. The raw data set is prepared for classification algorithms during the data pre-processing stage. The data is re-sampled with SMOTE according to the minority class during the data balancing step, and an over-sampling operation is carried out. To improve the classification process, we employed RFE to choose the best predictive features.

At this phase, the classification lithology system's performance is improved while the data set is balanced using a combination of SMOTE and RFE approaches. Finally, the CNN-RNN hybrid deep learning approach is used to classify various types of lithology during the classification step. The suggested strategy results in a considerable improvement in classification accuracy for all types of lithologies employed in this highly unbalanced data.

4. RESULT AND DISCUSSION

4.1 RESULT OF LITHOLOGY DATASET

The results of this investigation are presented and discussed in this part from two perspectives. The first is the effectiveness of the suggested method's CNN-RNN hybrid optimization classification algorithm. Second, the outcomes of the proposed method's balancing and RFE with four lithology datasets. In addition, the suggested CNN-RNN hybrid method was compared to a regular deep learning algorithm to demonstrate its efficacy with multi-imbalanced lithology datasets.

Four lithology datasets, containing multiple unbalanced difficulties and numerous non-significant characteristics, were employed in this investigation. These issues have the potential to skew classification and distort the outcomes. A new approach based on a combination of techniques was used to train various deep learning models for categorization. At first, to maximise the model's statistical ability, stage two preprocessing features with high values that are missing are removed. SMOTE was used to resample four lithography data sets, and the technique resulted in equal balance across all datasets. The total number of characteristics in the GP, GP1, and GP2 datasets is 18, whereas the GP3 dataset contains eight. RFE is used to eliminate the least effect features and obtain the best classification features. The scaling of data is the last phase in the processing method that converts the numerical information into a specified range. The data were divided to two sets:70% to be trained use and 30% for experimental use to evaluate the efficacy of DLMs.

In this work, we improved the classification procedure by creating and assessing four multi-balanced lithic datasets using deep learning models. ANN; RNN; CNN; and CNN-RNN modelling were constructed and refined based on the tests, and the optimal results were achieved by tuning the hyper-parameter using Grid Search CV. GP, GP1 and GP2 datasets yielded the best results after the grid-based search optimisation method was fine-tuned for all data. The epoch counts for the GP3 data were 200 and 250, respectively, the Drop Rate was 0.10 for all datasets, and the Convolution Layers were set to 128-for both-layers. Additionally, SoftMax: selected as the best being activated, Adam is selected as the most effective optimisation method for all the lithologies datasets, and the kernel is set to 3. Each neural network model's a hyperparameter, as show in Table 4.1.

Table 4.1: Hyper-Parameters of All Deep Learning Models for Lithology Datasets.

Datasets	Number of Layers	Number of Neurons	Activation Function	Loss Function	Epochs	Batch Size	Dropout
GP (ANN)	3	128,64,32	SoftMax	Categorical Crossentropy	100	45	-
GP (CNN)	2	128,128	SoftMax	Categorical Crossentropy	80	125	0.10
GP (RNN)	3	128,64,32	SoftMax	Categorical Crossentropy	50	185	0.10
GP (CNN-RNN)	2	128,128	SoftMax	Categorical Crossentropy	180	55	0.15
GP1 (ANN)	3	64,32,16	SoftMax	Categorical Crossentropy	40	55	-
GP1 (CNN)	2	128,128	SoftMax	Categorical Crossentropy	30	105	0.10
GP1 (RNN)	2	512, 256	SoftMax	Categorical Crossentropy	40	30	0.10
GP1 (CNN-RNN)	2	128,128	SoftMax	Categorical Crossentropy	210	185	0.10
GP2 (ANN)	3	64,32,16	SoftMax	Categorical Crossentropy	80	120	-
GP2 (CNN)	2	128,128	SoftMax	Categorical Crossentropy	30	55	0.10
GP2 (RNN)	3	128,64,32	SoftMax	Categorical Crossentropy	50	180	0.10
GP2 (CNN-RNN)	2	128,128	SoftMax	Categorical Crossentropy	150	152	0.10

Table 4.1: Hyper-Parameters of All DL Models for Lithology Datasets "Table Continued".

GP3 (ANN)	3	512,128,64,32	SoftMax	Categorical Crossentropy	150	115	-
GP3 (CNN)	2	128,128	SoftMax	Categorical Crossentropy	80	150	0.15
GP3 (RNN)	2	256,128	SoftMax	Categorical Crossentropy	100	120	0.10
GP3 (CNN-RNN)	2	128,128	SoftMax	Categorical Crossentropy	250	88	0.25

GP, GP1, GP2, & GP3 datasets yield the highest accuracy when using the hyper CNN-RNN, as Table 4.2 illustrates. Additionally, the categorization results from RNN, ANN, and CNN are good on their own. With the CNN approach, ANN, and CNN-RNN, the correctness ranges are (89% to 92%), (87% to 89%), and (91% to 93%) for the hyper approach, which performs better across the board among all measures of evaluation.

Table 4.2: Deep Learning Models Findings for Lithology Datasets.

Dataset	Method	Accuracy	Precision	F1-Score	Recall	G-Mean
GP	ANN	0.89	0.90	0.90	0.88	0.90
	RNN	0.89	0.90	0.88	0.87	0.90
	CNN	0.92	0.93	0.92	0.91	0.92
	CNN-RNN	0.93	0.93	0.93	0.92	0.94
GP1	ANN	0.88	0.88	0.88	0.87	0.88
	RNN	0.89	0.90	0.89	0.89	0.90

Table 4.2: Deep Learning Models Findings for Lithology Datasets "Table Continued".

	CNN	0.90	0.91	0.88	0.87	0.92
	CNN-RNN	0.92	0.92	0.92	0.91	0.93
GP2	ANN	0.88	0.89	0.87	0.87	0.89
	RNN	0.89	0.90	0.88	0.87	0.91
	CNN	0.90	0.91	0.90	0.90	0.93
	CNN-RNN	0.92	0.92	0.91	0.91	0.93
GP3	ANN	0.87	0.88	0.87	0.86	0.90
	RNN	0.88	0.88	0.88	0.87	0.91
	CNN	0.89	0.90	0.89	0.88	0.92
	CNN-RNN	0.91	0.91	0.91	0.90	0.93

These assessment measurements, such accuracy, are unworkable in the method of classification as the data set includes the issue of several unbalanced classes. If not even one minority class is correctly predicted, the prediction may become skewed in favour of the larger class and misleadingly show greater accuracy [65]. For this reason, unbalanced datasets were tested using the F1-Score and G-Mean evaluation measures. Considering that the G-Mean combines both sensitivity and specificity, while the F1-Score combines the precision and recall. Table 4.2 indicates that all our models now have F1-Score and G-Mean that indicate well-balanced processing performance. Moreover, (AUC) is employed since it includes two metrics for measuring performance, namely (TPR) and (FPR), which are used concurrently [66].

The hardest part of any ML method is generalising the model to the point where it can forecast appropriate outcomes using fresh data instead of just the data it was trained on. Comparing Train accuracy vs. accuracy of test over several epochs is a helpful method to confirm whether the model has been properly trained. This is required to prevent the model from getting too or too trained, which would reduce its capacity to predict accurately by

causing it to start memorising the training set. As seen by Figures 4.1–4.3, three ML analysis techniques are employed to evaluate the effectiveness of our suggested model.

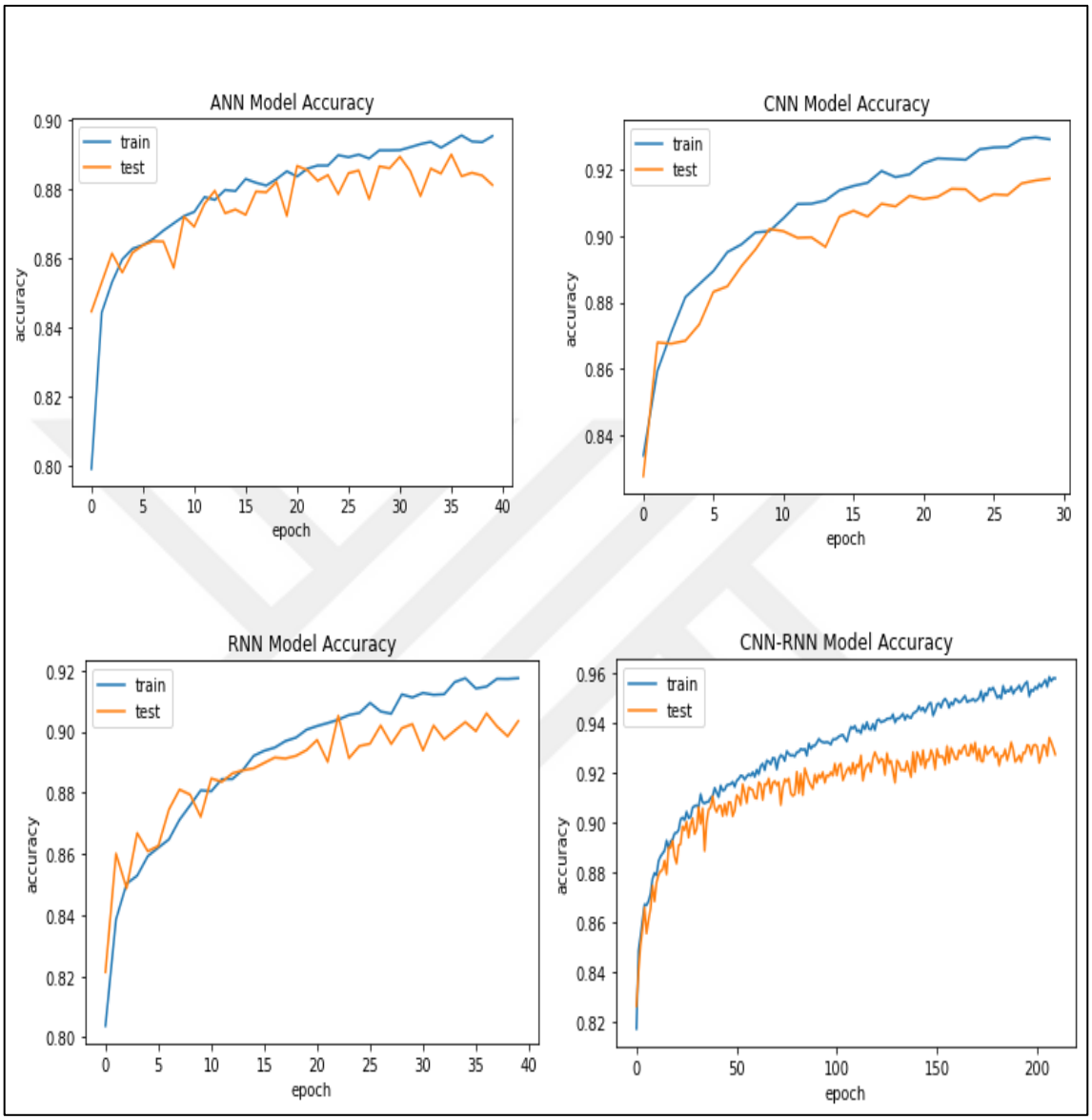

Figure 4.1: Analysis of Training Accuracy vs Test Accuracy of GP Dataset.

In Figure 4.1, for determining whether DLMs are properly trained and have not overfitted over multiple epochs, training accuracy is compared to test accuracy. The blue line indicates the Training Accuracy, while the orange line indicates the Test Accuracy. The DL models show good compatibility of all models with the GP group dataset, because there is no substantial difference between both the accuracy of Train-data and accuracy of Test-data, indicating that all models are a good match and do not have overfitting or underfitting issues.

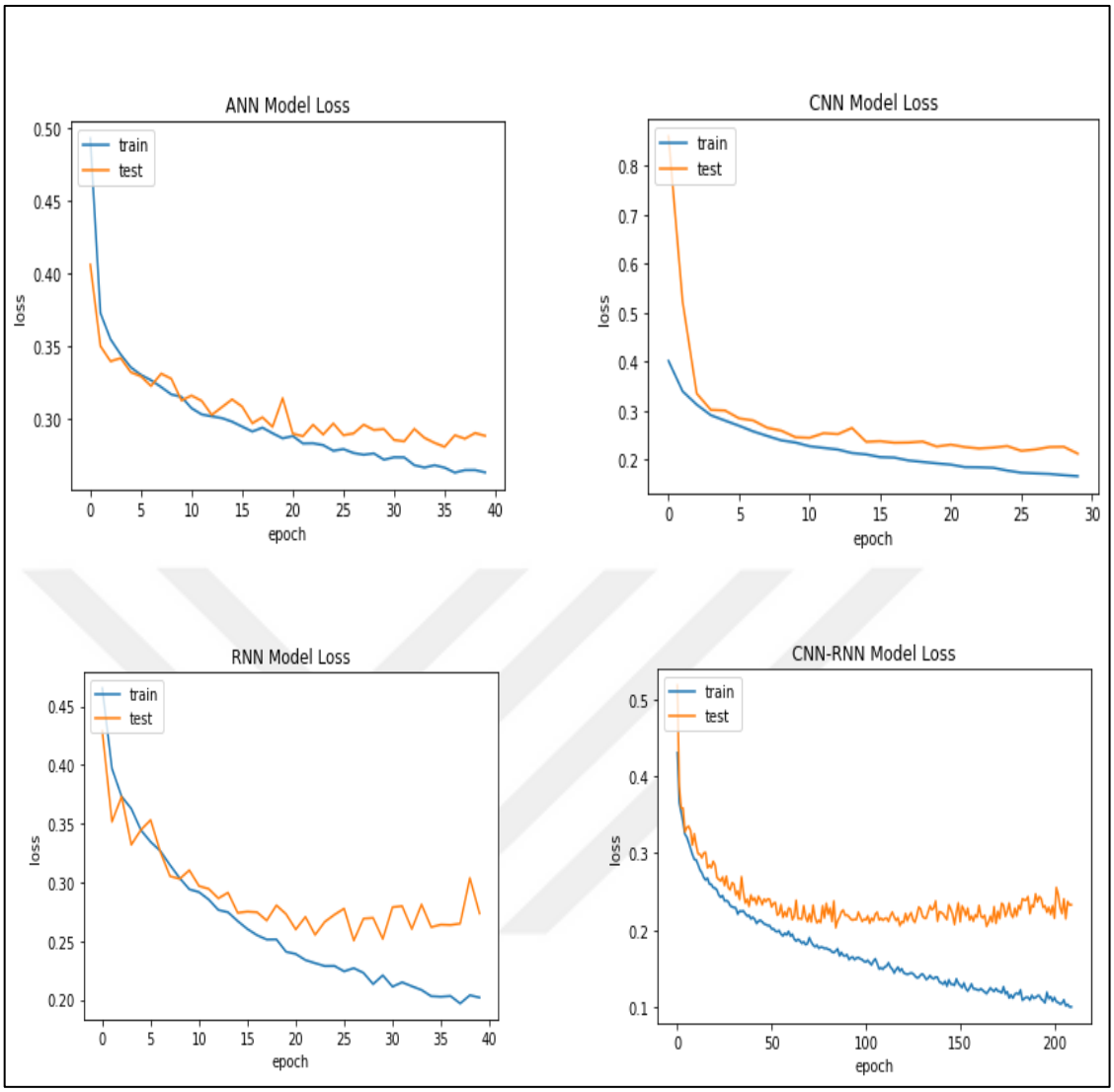

Figure 4.2: Analysis of Training Loss vs Test Loss Accuracy of GP Dataset.

In Figure 4.2, the efficacy of each DLM was assessed using training loss against test loss on the GP dataset. Across multiple epochs, the test loss is represented by the colour orange and the training loss by the blue line. To minimise data loss, all DLMs have modified dropout with an additional parameter.

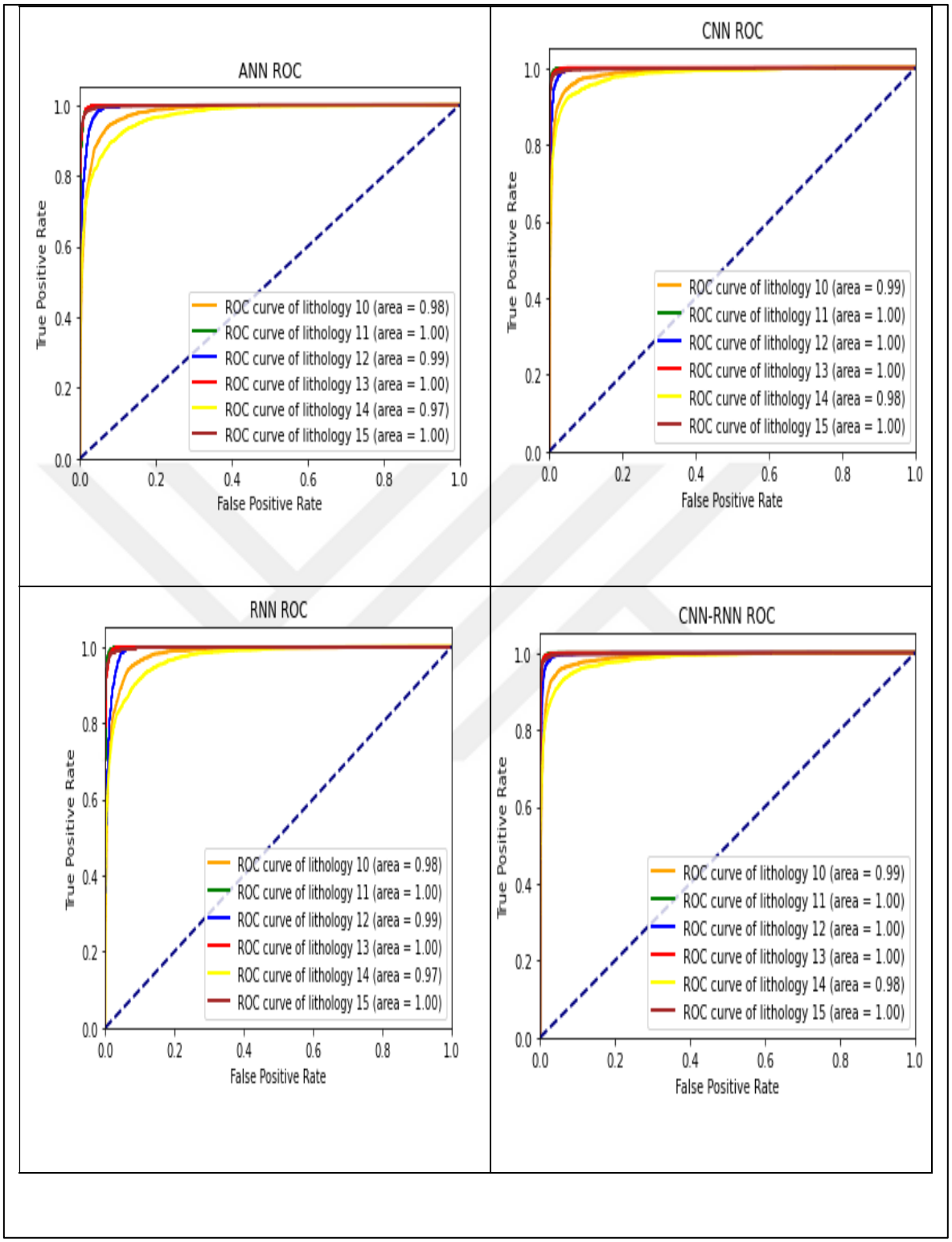

Figure 4.3: ROC Curve Analysis for GP Dataset.

In Figure 4.3, DL models utilised in the GP six lithology classes' ROC curves. GP group's lithology via coding 10 is represented by the orange-coloured curve, lithology via coding 11 by the green curve, lithology via coding 12 by the blue curve, lithology with code 13 by the coloured red curve, lithology via coding 14 by the yellow-coloured curve, and lithology via coding 15 by the brown-coloured curve.

Table 4.3, describe the suggested study comparisons using the lithology dataset that is unbalanced across multiple classes. Bressan et al. [25] Use 4 datasets to classify lithology. The 4 datasets were trained and tested with 5-fold cross-validation by Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Multilayer Perceptron (MLP). The best accuracy obtained using MLP in the GP2 data set is 85% accuracy.

Table 4.3: The Comparison of This Research with the Earlier Research.

Author	Dataset	Accuracy	Method
Bressan et al. [33]	GP	0.83	ML
	GP1	0.73	ML
	GP2	0.85	ML
	GP3	0.80	ML
This Work	GP	0.93	DL
	GP1	0.92	DL
	GP2	0.92	DL
	GP3	0.91	DL

4.2 RESULT OF CONVOLUTIONAL DATASET

To assess the efficacy of the suggested hyper-DL model, three UCI data were employed. Tables 4.5–4.6 display all the hyperparameters for DL-models that were adjusted by Grid-search optimisation for: Thyroid Disease, Car Evaluation, and Dermatology datasets.

Table 4.4: Car Evaluation Dataset Hyper-Parameters.

Method	Number of Layers	Number of Neurons	Activation Function	Loss Function	Epochs	Batch Size	Dropout
ANN	3	64,32,16	SoftMax	Categorical Crossentropy	30	15	-
CNN	2	128,128	SoftMax	Categorical Crossentropy	40	55	0.25
RNN	2	512,256	SoftMax	Categorical Crossentropy	40	15	0.10
CNN-RNN	3	128, 128, 128	SoftMax	Categorical Crossentropy	50	16	0.20

Table 4.5: Thyroid-Disease Dataset Hyper-Parameters.

Method	Number of Layers	Number of Neurons	Activation Function	Loss Function	Epochs	Batch Size	Dropout
ANN	3	64,32,16	SoftMax	Categorical Crossentropy	40	55	-
CNN	2	64,64	SoftMax	Categorical Crossentropy	30	100	0.10
RNN	2	512,256	SoftMax	Categorical Crossentropy	40	55	0.10
CNN-RNN	3	128, 128	SoftMax	Categorical Crossentropy	50	220	0.15

Table 4.6: Dermatology Dataset Hyper-Parameters.

Method	Number of Layers	Number of Neurons	Activation Function	Loss Function	Epochs	Batch Size	Dropout
ANN	3	64,32,16	SoftMax	Categorical Crossentropy	30	5	-
CNN	2	64, 64	SoftMax	Categorical Crossentropy	30	50	0.10
RNN	2	512, 256	SoftMax	Categorical Crossentropy	40	30	0.10
CNN-RNN	2	128,128	SoftMax	Categorical Crossentropy	40	11	0.10

As shown in Table 4.7-4.9 the proposed hyper deep learning model (CNN-RNN) show a high evaluation result for all use multiclass datasets. In contrast all other deep learning models show a high performance with accuracy range from 96% to 99%, also all models show significant improvement with G-Mean with range from 97% to 100%.

Table 4.7: Result of Car Evaluation Dataset.

Method	Accuracy	Precision	F1-Score	Recall	G-Mean
ANN	98.1%	99.4%	97.3%	97%	99%
RNN	99.1%	98.3%	99.1%	99.1%	99%
CNN	99%	99.1%	95%	95.3%	99%
CNN-RNN	99.2%	99.2%	99.1%	99.3%	100%

Table 4.8: Result of Thyroid-Disease Dataset.

Method	Accuracy	Precision	F1-Score	Recall	G-Mean
ANN	98%	98%	98%	98%	98%
RNN	99%	98%	99%	98%	98%
CNN	99%	99%	99%	99%	99%
CNN-RNN	99%	99%	99%	99%	100%

Table 4.9: Result of Dermatology Dataset.

Method	Accuracy	Precision	F1-Score	Recall	G-Mean
ANN	98.1%	95.2%	94%	93.3%	98%
RNN	96.2%	98.1%	92%	90.2%	97%
CNN	99.1%	99.1%	99.2%	99.1%	99%
CNN-RNN	99.24%	99.2%	99.3%	99.4%	99.20%

The results of different analyses were used such as: Training Accuracy versus Test Accuracy, and Training Loss vs Testing Loss, and ROC curve to evaluate the performance of all convolutional datasets. The first analysis report for convolutional data shown in Figure 4.4 – 4.6, the second convolutional data analysis shown in Figure 4.7 – 4.9, and the third data shown in Figure 4.10 – 4.12.

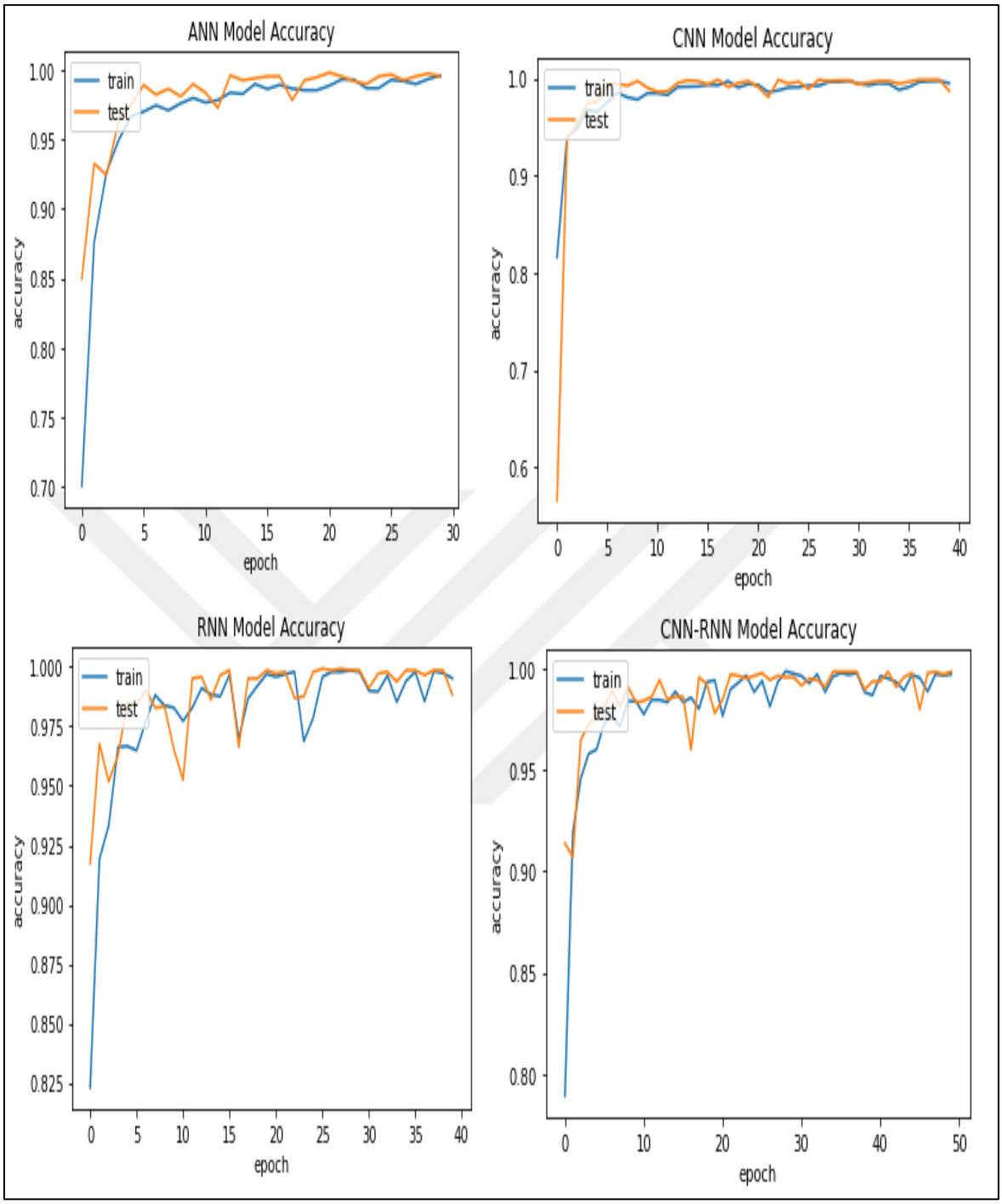

Figure 4.4: Training Accuracy vs Test Accuracy Analysis for Car Evaluation Data.

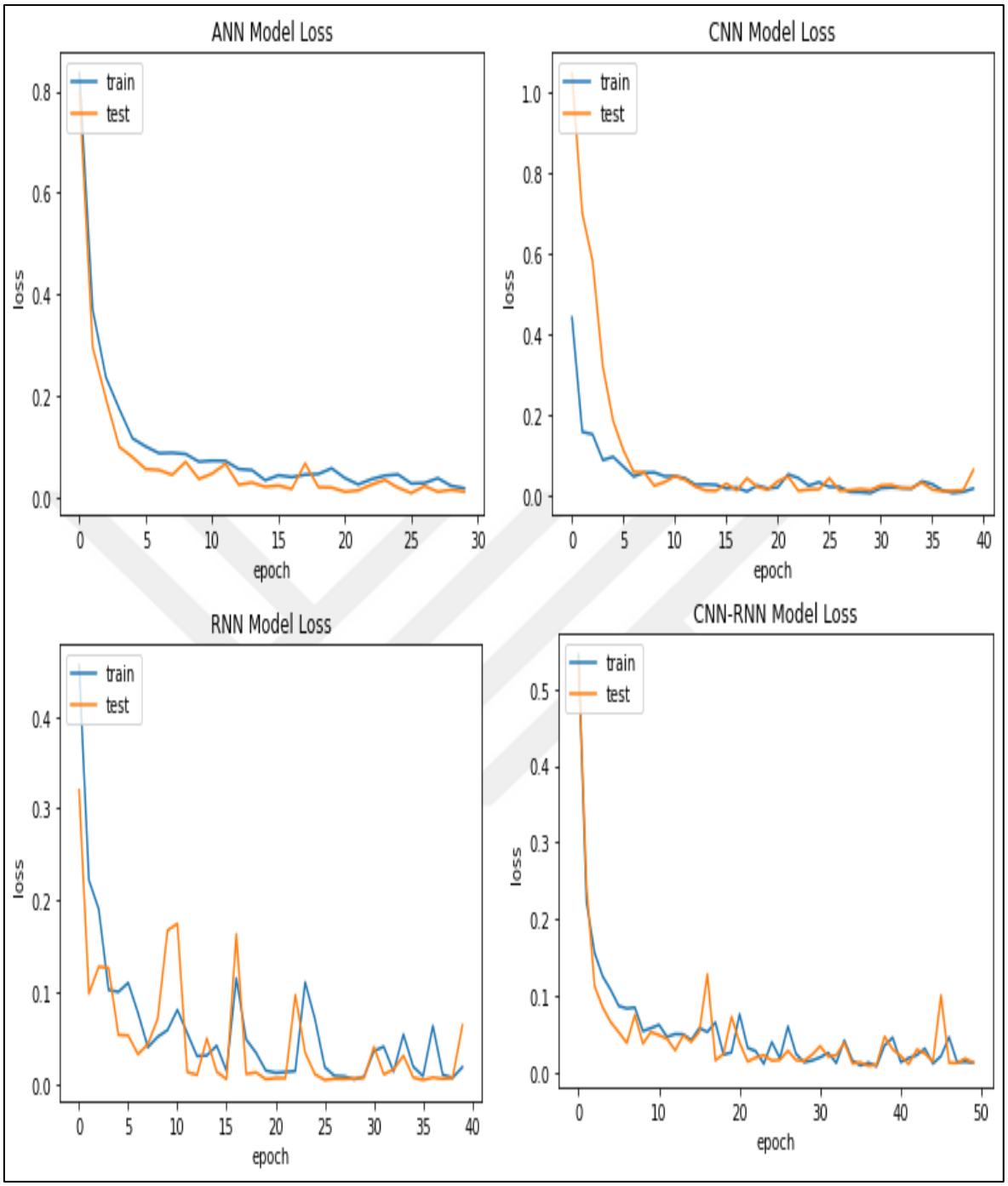

Figure 4.5: Training Loss vs Test Loss for Car Evaluation Data.

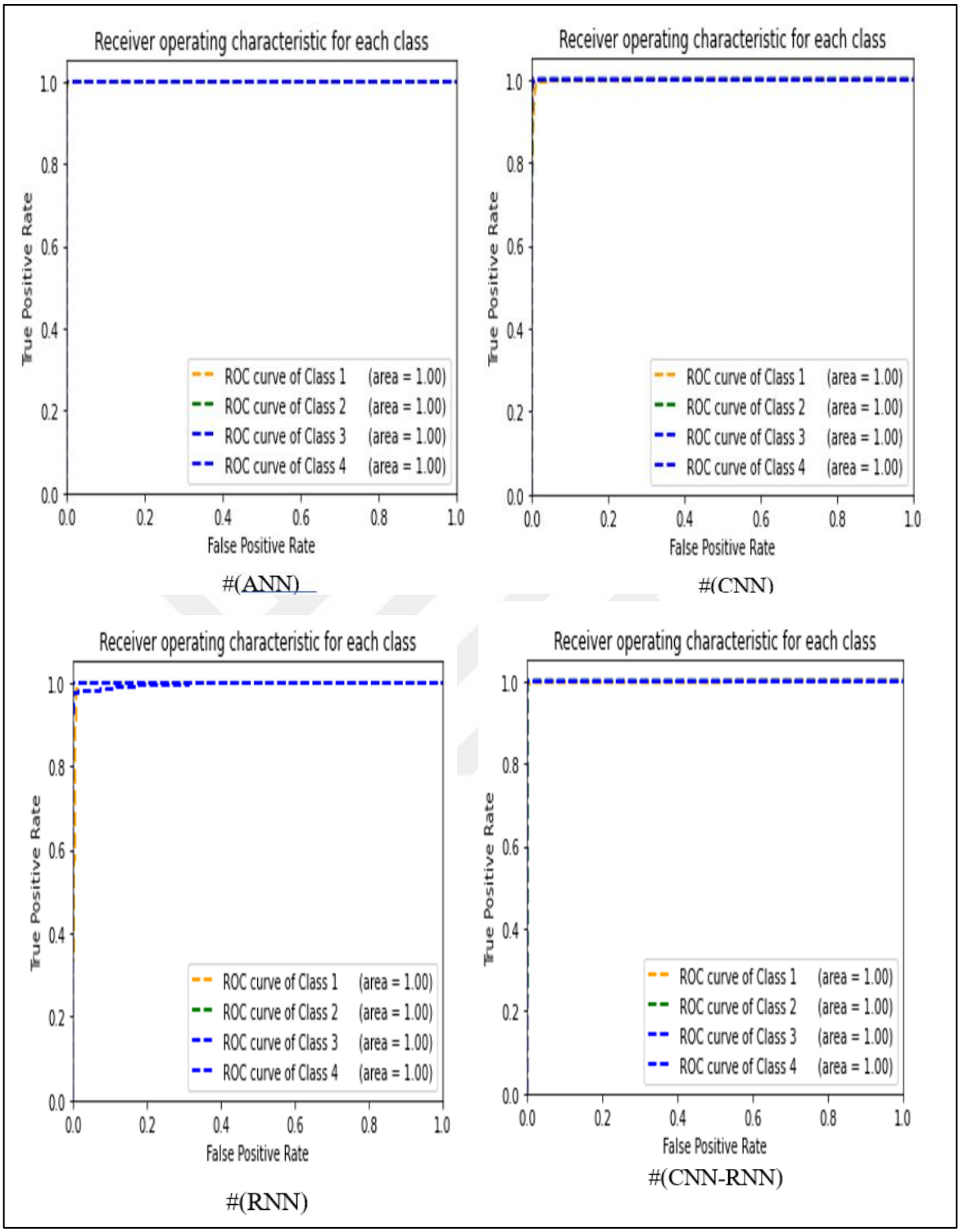

Figure 4.6: ROC Curve Analysis for Car Evaluation Data.

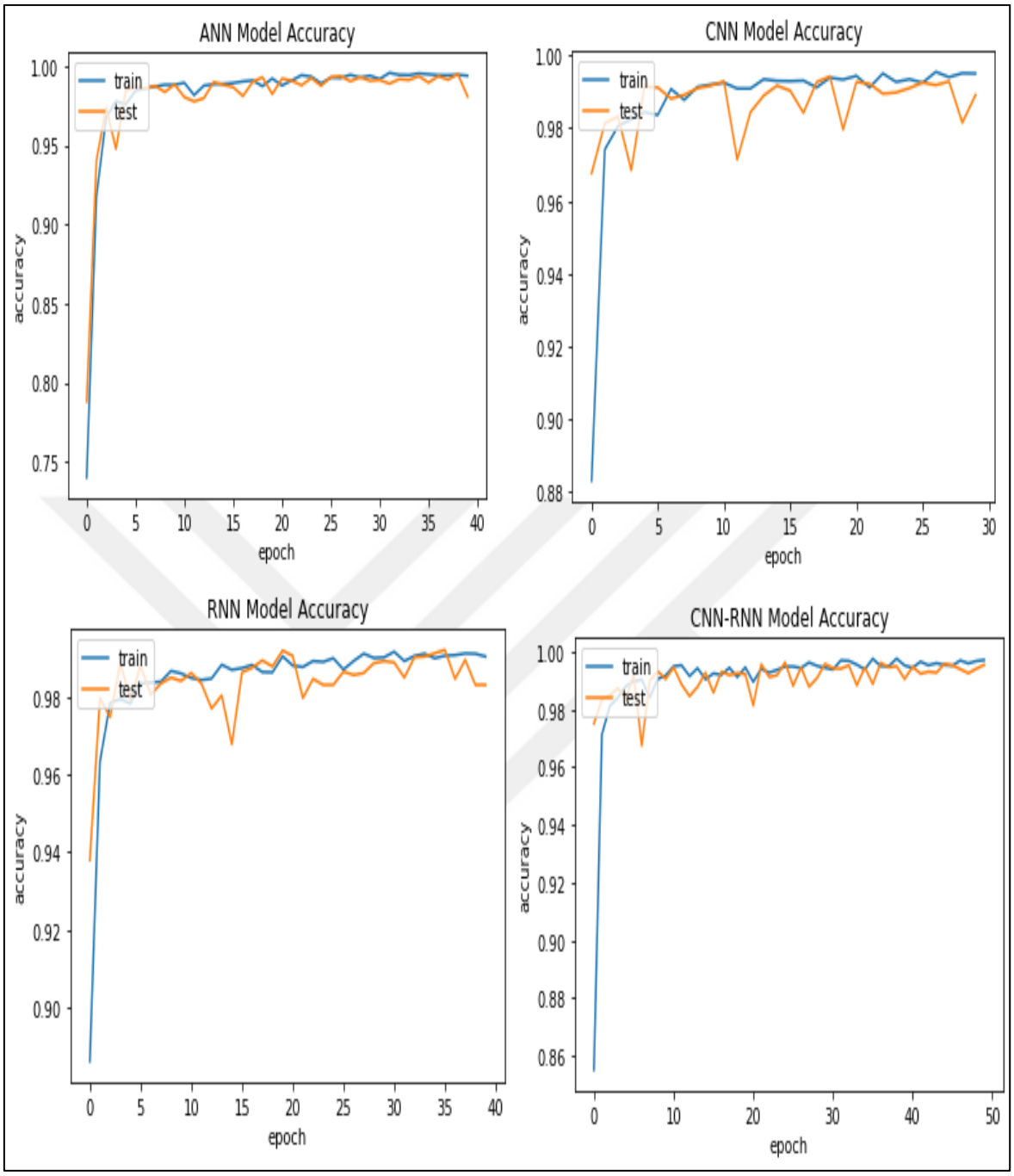

Figure 4.7: Training Accuracy vs Test Accuracy Analysis for Thyroid-Disease Data.

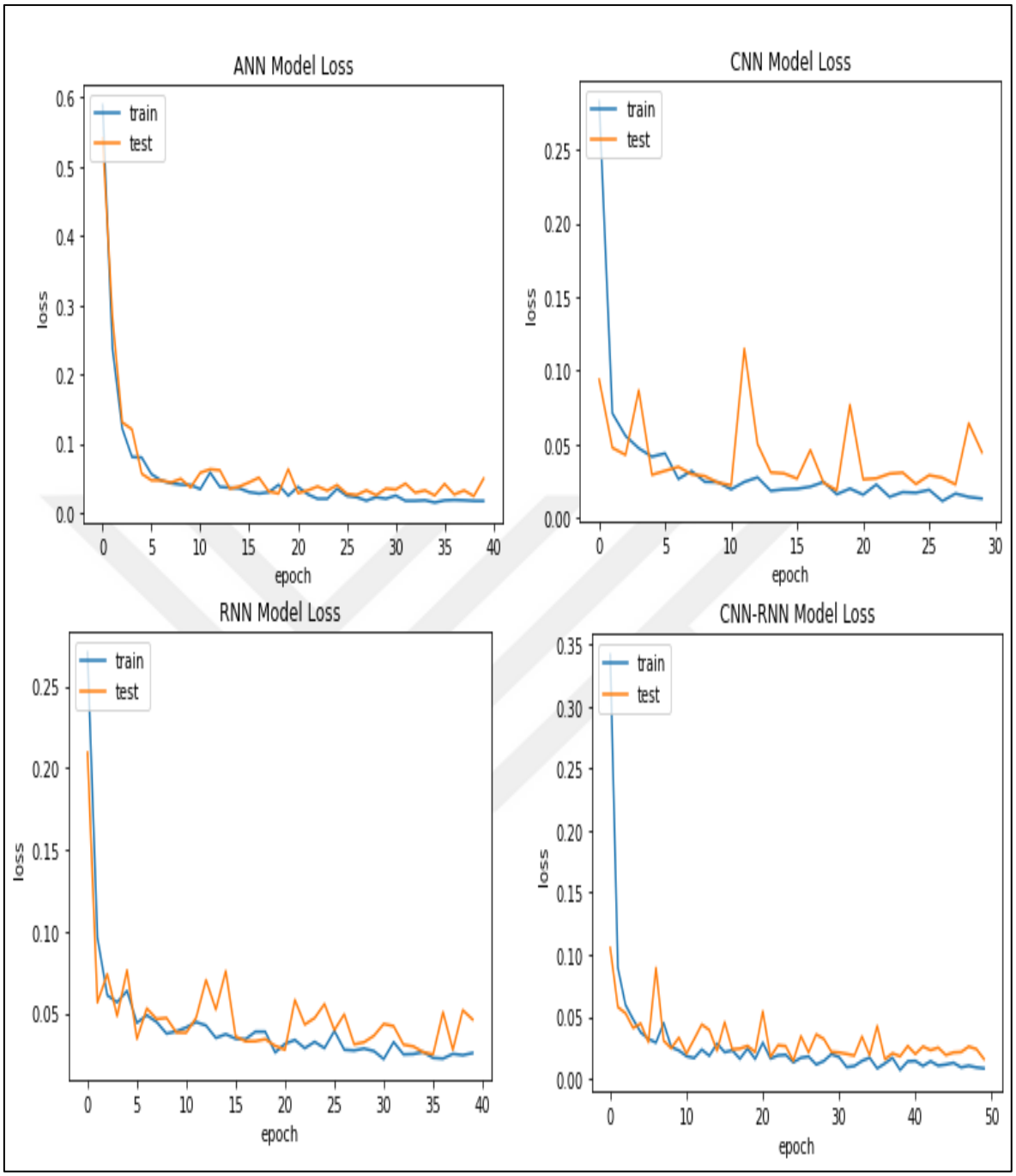

Figure 4.8: Training Loss vs Test Loss for Thyroid-Disease Data.

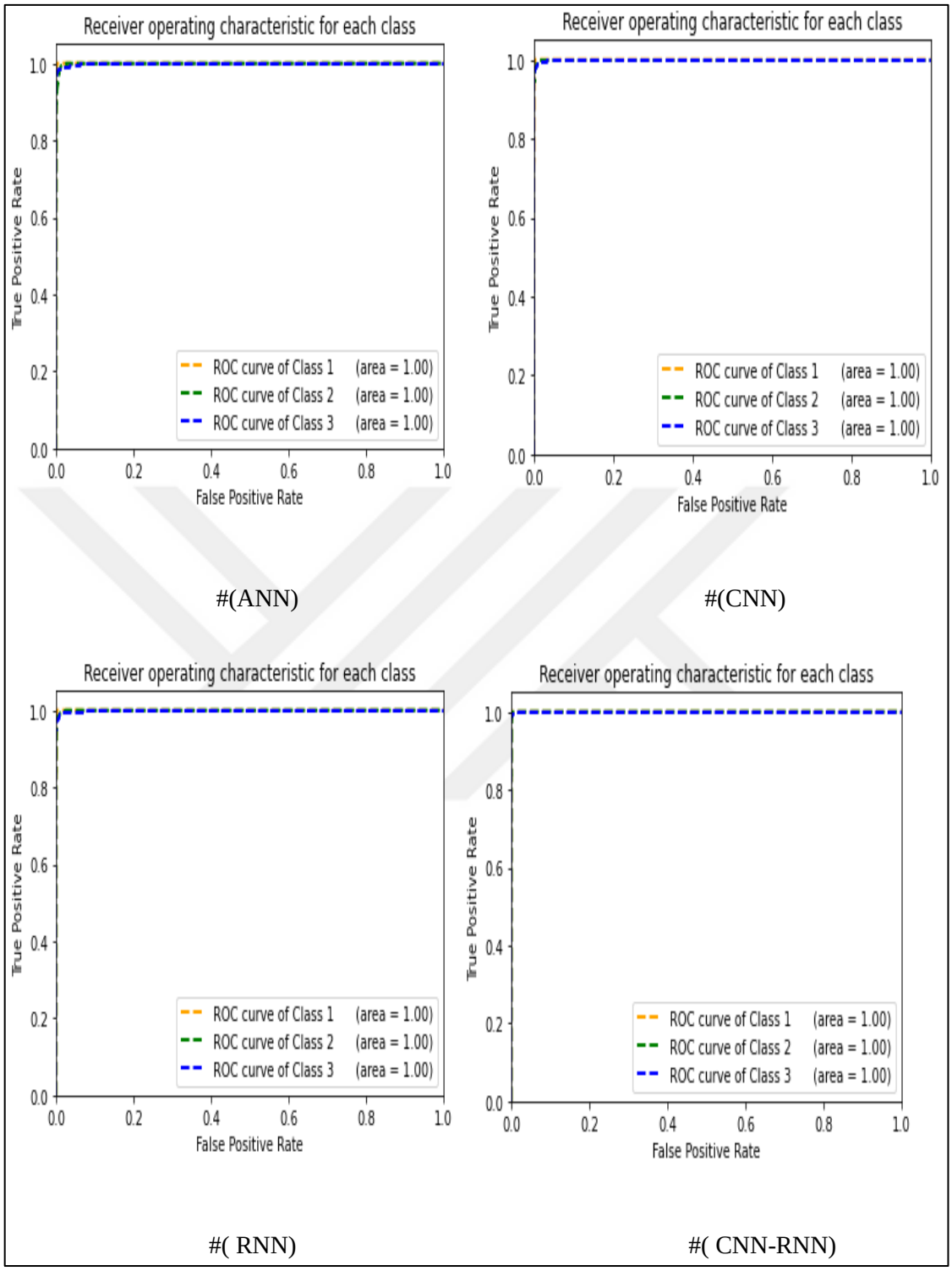

Figure 4.9: ROC Curve Analysis for Thyroid-Disease Data.

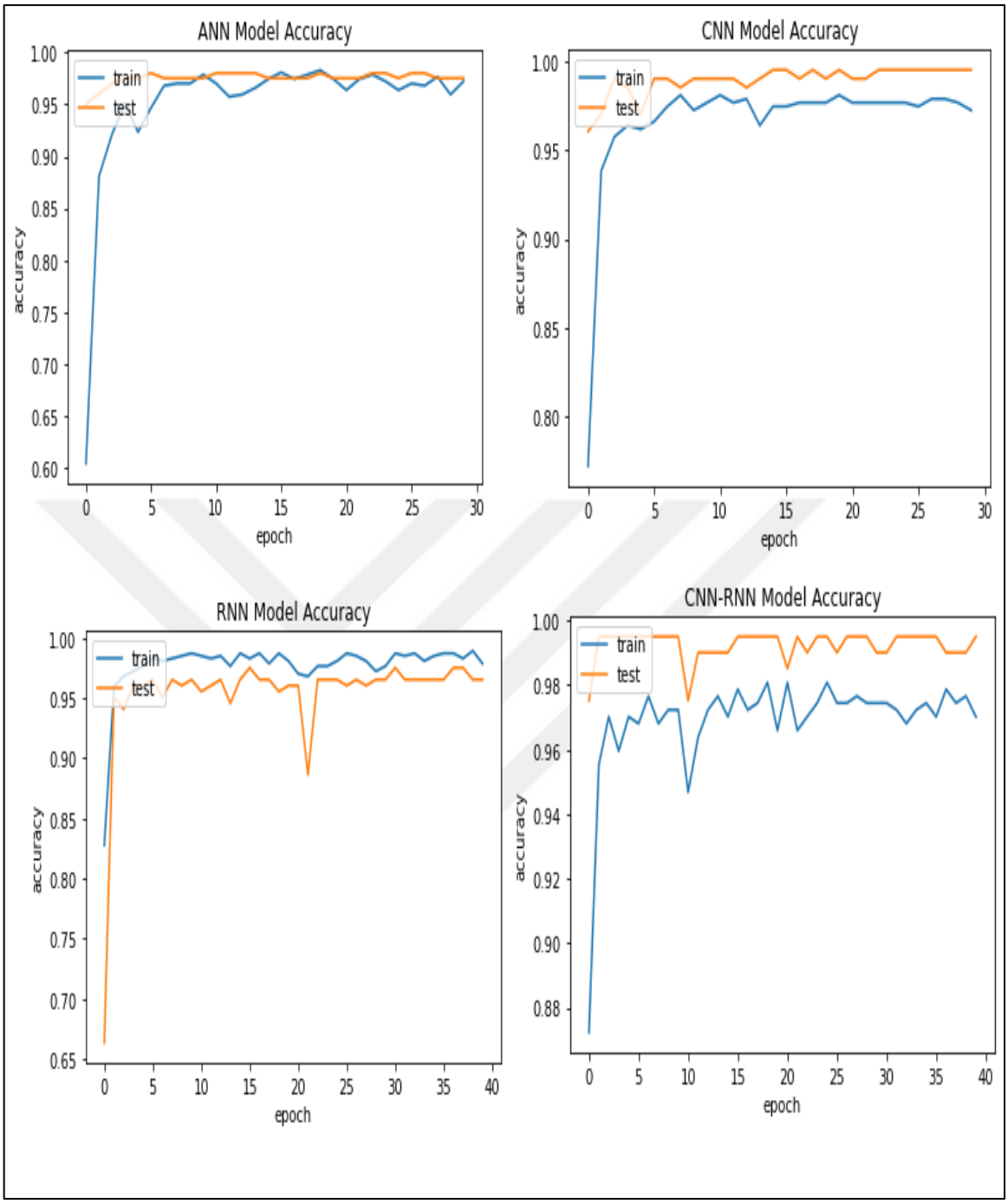

Figure 4.10: Training Accuracy vs Test Accuracy Analysis for Dermatology Data.

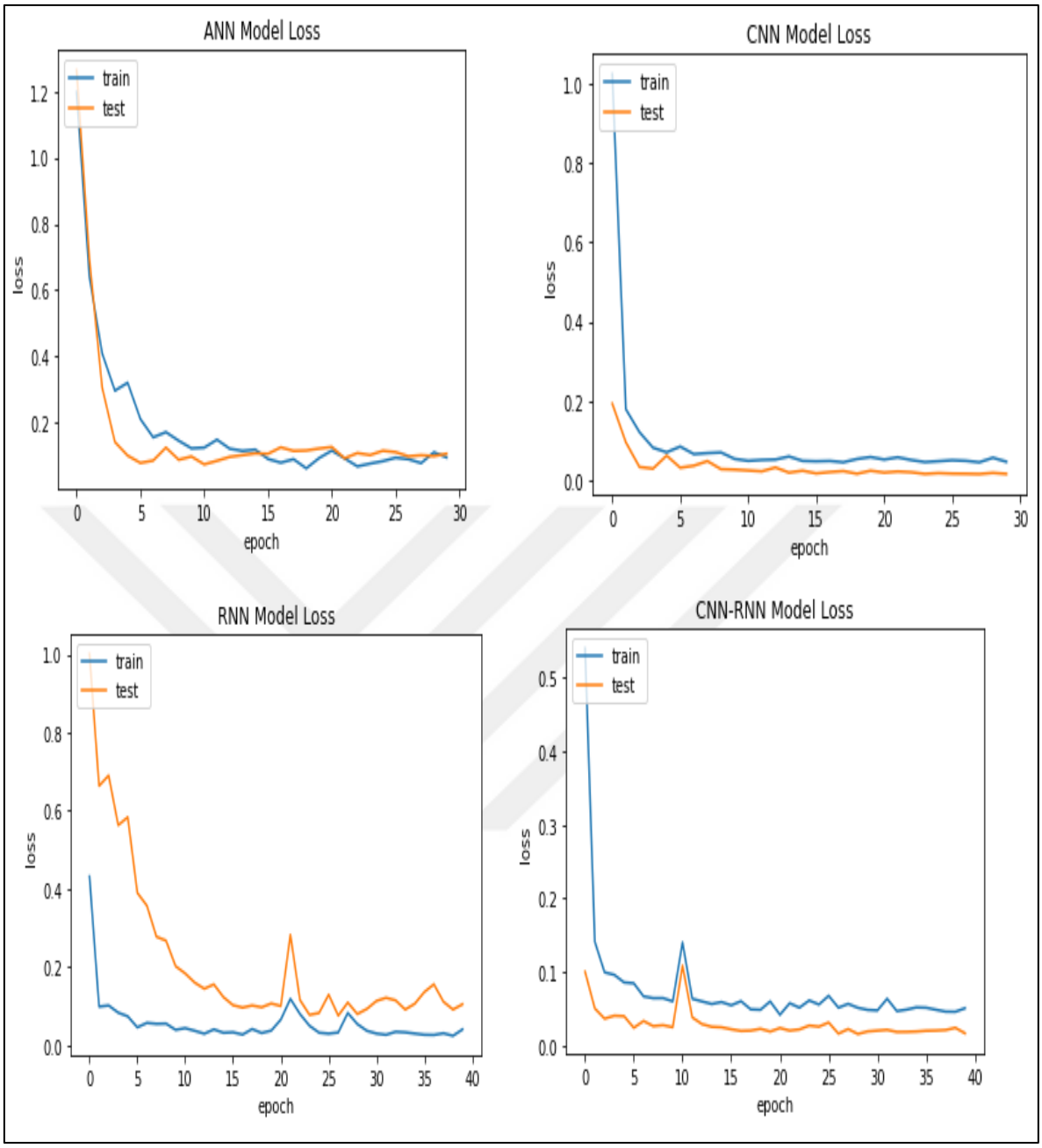

Figure 4.11: Training Loss vs Test Loss for Dermatology Data.

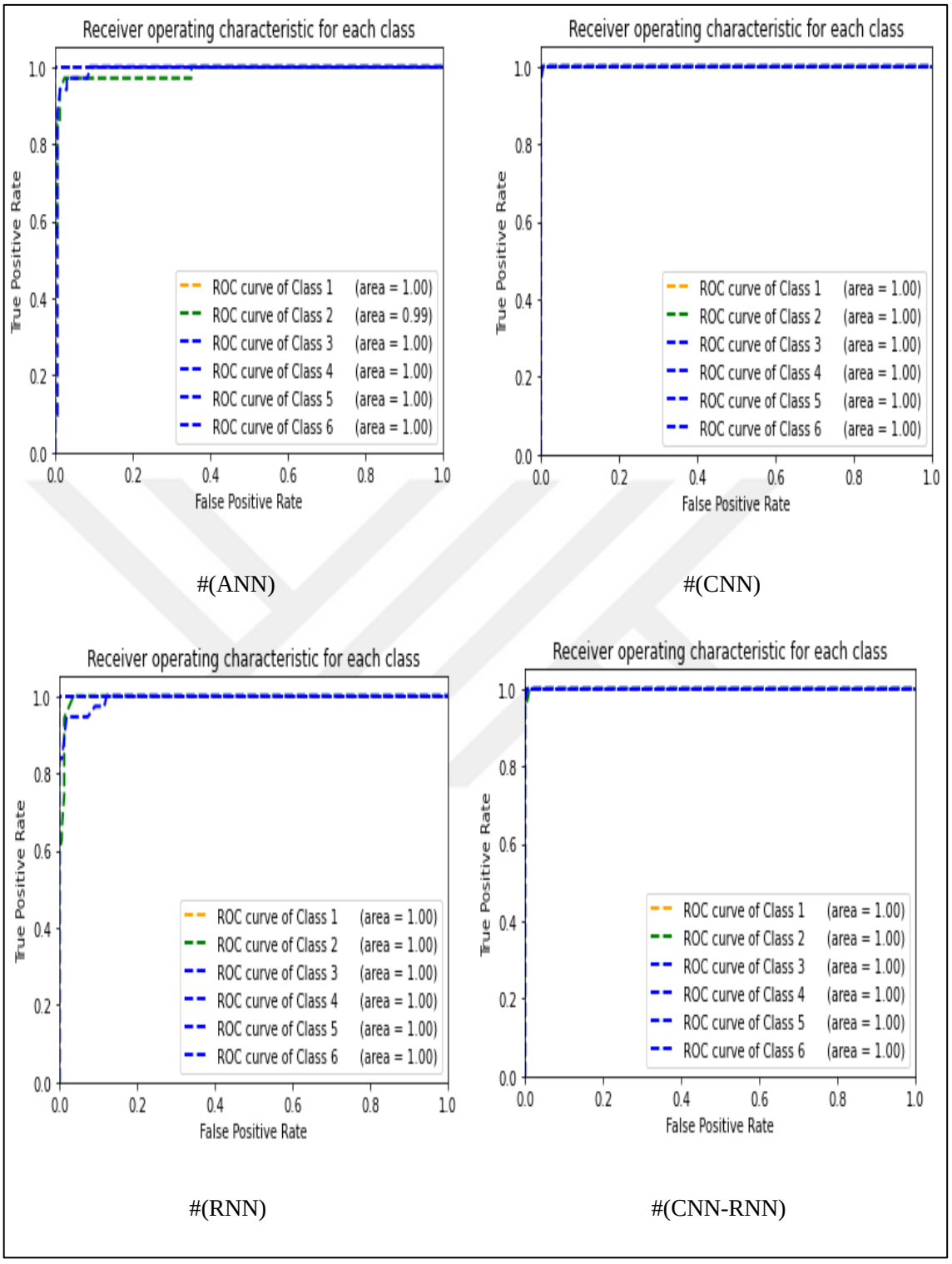

Figure 4.12: ROC Curve Analysis for Dermatology Data.

5. CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

This research presents an advanced hybrid DL model constructed on unbalanced multi-class lithology data sets. Selecting wrapper features and information balances are integrated. Combination technique (SMOTE + RFE) is used to perform data balance and to determine optimal features for classification. Furthermore, the lithology classification process employed consisted of an CNN-RNN DL model, that blends CNN and RNN models. The UCI offered three multi-class imbalanced datasets and four highly dimensional and multi-class lithography datasets. To increase overall performance, all model hyperparameters were fine-tuned using a grid search optimisation technique. The results indicate that compared to classical DLMs, CNN and RNN perform superior in multiple classes lithography for classification. The RNN uses the identical capabilities for each input, while CNN successfully pulls attributes from the lithography data. The outcome of that input is reliant on the prior computing through its own inbuilt memory. The efficacy of the incidents was confirmed using Accuracy, Precision, F1-Score, Recall, G-Mean, and ROC curves. A consequence of experimental findings, proposed method outperformed other methods, with an accuracy of 93% and a ROC curve of 95%.

5.2 FUTURE WORK

Future works will focus on the following areas:

a. It might create a hyper oversampling method for working with binary and multi class data.

b. Extending the recommended technique to other classification systems, such as medical diagnosis, with a corresponding change in the dataset used.

c. Apply ensemble learning kernels with oversampling method.

REFERENCES

[1] Y. Xiong and R. Zuo, “Robust Feature Extraction for Geochemical Anomaly Recognition Using a Stacked Convolutional Denoising Autoencoder,” Math. Geosci. 2021, pp. 1–22, Feb. 2021, doi: 10.1007/S11004-021-09935-Z.

[2] R. Zuo, Y. Xiong, J. Wang, and E. J. M. Carranza, “Deep learning and its application in geochemical mapping,” Earth-Science Rev., vol. 192, pp. 1–14, May 2019, doi: 10.1016/J.EARSCIREV.2019.02.023.

[3] M. Reichstein et al., “Deep learning and process understanding for data-driven Earth system science,” Nat. 2019 5667743, vol. 566, no. 7743, pp. 195–204, Feb. 2019, doi: 10.1038/s41586-019-0912-1.

[4] Y. Xiong and R. Zuo, “Recognizing multivariate geochemical anomalies for mineral exploration by combining deep learning and one-class support vector machine,” Comput. Geosci., vol. 140, p. 104484, Jul. 2020, doi: 10.1016/J.CAGEO.2020.104484.

[5] T. Li, R. Zuo, Y. Xiong, and Y. Peng, “Random-Drop Data Augmentation of Deep Convolutional Neural Network for Mineral Prospectivity Mapping,” Nat. Resour. Res. 2020 301, vol. 30, no. 1, pp. 27–38, Sep. 2020, doi: 10.1007/S11053-020-09742-Z.

[6] A. Brcković, M. Kovačević, M. Cvetković, I. Kolenković Močilac, D. Rukavina, and B. Saftić, “Application of artificial neural networks for lithofacies determination based on limited well data,” Cent. Eur. Geol., vol. 60, no. 3, pp. 299–315, Dec. 2017, doi: 10.1556/24.60.2017.012.

[7] R. Latifovic, D. Pouliot, and J. Campbell, “Assessment of Convolution Neural Networks for Surficial Geology Mapping in the South Rae Geological Region, Northwest Territories, Canada,” Remote Sens. 2018, Vol. 10, Page 307, vol. 10, no. 2, p. 307, Feb. 2018, doi: 10.3390/RS10020307.

[8] A. E. Bruno et al., “Classification of crystallization outcomes using deep convolutional neural networks,” PLoS One, vol. 13, no. 6, p. e0198883, Jun. 2018,

doi: 10.1371/JOURNAL.PONE.0198883.

[9] T. Kumar, S. N. Kumar, and G. S. Rao, “Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India,” J. Appl. Geophys., p. 104605, Mar. 2022, doi: 10.1016/J.JAPPGEO.2022.104605.

[10] R. Kiran, P. Dansena, S. Salehi, and V. K. Rajak, “Application of machine learning and well log attributes in geothermal drilling,” Geothermics, vol. 101, p. 102355, May 2022, doi: 10.1016/J.GEOTHERMICS.2022.102355.

[11] I. Serbouti, M. Raji, and M. Hakdaoui, “Lithological Mapping for a Semi-arid Area Using GEOBIA and PBIA Machine Learning Approaches with Sentinel-2 Imagery: Case Study of Skhour Rehamna, Morocco,” Adv. Sci. Technol. Innov., pp. 143–156, 2022, doi: 10.1007/978-3-030-80458-9_11.

[12] Y. Dai, M. Khandelwal, Y. Qiu, J. Zhou, M. Monjezi, and P. Yang, “A hybrid metaheuristic approach using random forest and particle swarm optimization to study and evaluate backbreak in open-pit blasting,” Neural Comput. Appl. 2021 348, vol. 34, no. 8, pp. 6273–6288, Jan. 2022, doi: 10.1007/S00521-021-06776-Z.

[13] H. Fathipour-Azar, “Polyaxial Rock Failure Criteria: Insights from Explainable and Interpretable Data-Driven Models,” Rock Mech. Rock Eng. 2022, pp. 1–19, Feb. 2022, doi: 10.1007/S00603-021-02758-8.

[14] S. Lin, H. Zheng, B. Han, Y. Li, C. Han, and W. Li, “Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction,” Acta Geotech., pp. 1–26, Jan. 2022, doi: 10.1007/S11440-021-01440-1/TABLES/3.

[15] I. Mondal and K. H. Singh, “Core-log integration and application of machine learning technique for better reservoir characterisation of Eocene carbonates, Indian offshore,” Energy Geosci., vol. 3, no. 1, pp. 49–62, Jan. 2022, doi: 10.1016/J.ENGEOS.2021.10.006.

[16] B. Ullah, M. Kamran, and Y. Rui, “Predictive Modeling of Short-Term Rockburst for the Stability of Subsurface Structures Using Machine Learning Approaches: t-SNE,

K-Means Clustering and XGBoost," Math. 2022, Vol. 10, Page 449, vol. 10, no. 3, p. 449, Jan. 2022, doi: 10.3390/MATH10030449.

[17] S. Chauhan et al., "Processing of rock core microtomography images: Using seven different machine learning algorithms," Comput. Geosci., vol. 86, pp. 120–128, Jan. 2016, doi: 10.1016/J.CAGEO.2015.10.013.

[18] M. He, H. Gu, and H. Wan, "Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir," J. Pet. Sci. Eng., vol. 194, p. 107498, Nov. 2020, doi: 10.1016/J.PETROL.2020.107498.

[19] Y. Xie, C. Zhu, Y. Lu, and Z. Zhu, "Towards Optimization of Boosting Models for Formation Lithology Identification," Math. Probl. Eng., vol. 2019, 2019, doi: 10.1155/2019/5309852.

[20] Z. Ye, S. Guo, D. Chen, H. Wang, and S. Li, "Drilling formation perception by supervised learning: Model evaluation and parameter analysis," J. Nat. Gas Sci. Eng., vol. 90, p. 103923, Jun. 2021, doi: 10.1016/J.JNGSE.2021.103923.

[21] H. A. N. Ruiyi, W. A. N. G. Zhuwen, W. A. N. G. Wenhua, X. U. Fanghui, Q. I. Xinghua, and C. U. I. Yitong, "Lithology identification of igneous rocks based on XGboost and conventional logging curves, a case study of the eastern depression of Liaohe Basin," J. Appl. Geophys., vol. 195, p. 104480, Dec. 2021, doi: 10.1016/J.JAPPGEO.2021.104480.

[22] A. Alsaihati, S. Elkatatny, and H. Gamal, "Rate of penetration prediction while drilling vertical complex lithology using an ensemble learning model," J. Pet. Sci. Eng., vol. 208, p. 109335, Jan. 2022, doi: 10.1016/J.PETROL.2021.109335.

[23] Y. Zou, Y. Chen, and H. Deng, "Gradient Boosting Decision Tree for Lithology Identification with Well Logs: A Case Study of Zhaoxian Gold Deposit, Shandong Peninsula, China," Nat. Resour. Res. 2021 305, vol. 30, no. 5, pp. 3197–3217, Jun. 2021, doi: 10.1007/S11053-021-09894-6.

[24] C. M. Saporetti, L. G. Da Fonseca, and E. Pereira, "A Lithology Identification

Approach Based on Machine Learning with Evolutionary Parameter Tuning," IEEE Geosci. Remote Sens. Lett., vol. 16, no. 12, pp. 1819–1823, Dec. 2019, doi: 10.1109/LGRS.2019.2911473.

[25] G. Li, Y. Qiao, Y. Zheng, Y. Li, and W. Wu, "Semi-Supervised Learning Based on Generative Adversarial Network and Its Applied to Lithology Recognition," IEEE Access, vol. 7, pp. 67428–67437, 2019, doi: 10.1109/ACCESS.2019.2918366.

[26] P. X. Li, X. T. Feng, G. L. Feng, Y. X. Xiao, and B. R. Chen, "Rockburst and microseismic characteristics around lithological interfaces under different excavation directions in deep tunnels," Eng. Geol., vol. 260, p. 105209, Oct. 2019, doi: 10.1016/J.ENGGEO.2019.105209.

[27] Z. H. Xu, W. Y. Wang, P. Lin, L. C. Nie, J. Wu, and Z. M. Li, "Hard-rock TBM jamming subject to adverse geological conditions: Influencing factor, hazard mode and a case study of Gaoligongshan Tunnel," Tunn. Undergr. Sp. Technol., vol. 108, p. 103683, Feb. 2021, doi: 10.1016/J.TUST.2020.103683.

[28] V. A. Dev and M. R. Eden, "Evaluating the Boosting Approach to Machine Learning for Formation Lithology Classification," Comput. Aided Chem. Eng., vol. 44, pp. 1465–1470, Jan. 2018, doi: 10.1016/B978-0-444-64241-7.50239-1.

[29] J. Sun et al., "Optimization of models for a rapid identification of lithology while drilling - A win-win strategy based on machine learning," J. Pet. Sci. Eng., vol. 176, pp. 321–341, May 2019, doi: 10.1016/J.PETROL.2019.01.006.

[30] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," Comput. Chem. Eng., vol. 128, pp. 392–404, Sep. 2019, doi: 10.1016/J.COMPCHEMENG.2019.06.001.

[31] K. Maxwell, M. Rajabi, and J. Esterle, "Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques," Int. J. Coal Geol., vol. 214, p. 103284, Oct. 2019, doi: 10.1016/J.COAL.2019.103284.

[32] S. Shayeganpour, M. H. Tangestani, and P. V. Gorsevski, "Machine learning and multi-sensor data fusion for mapping lithology: A case study of Kowli-kosh area, SW

Iran," Adv. Sp. Res., vol. 68, no. 10, pp. 3992–4015, Nov. 2021, doi: 10.1016/J.ASR.2021.08.003.

[33] T. S. Bressan, M. Kehl de Souza, T. J. Girelli, and F. C. Junior, "Evaluation of machine learning methods for lithology classification using geophysical data," Comput. Geosci., vol. 139, p. 104475, Jun. 2020, doi: 10.1016/J.CAGEO.2020.104475.

[34] A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi, "A framework for feature selection through boosting," Expert Syst. Appl., vol. 187, p. 115895, Jan. 2022, doi: 10.1016/J.ESWA.2021.115895.

[35] J. Zhang, T. Wang, W. W. Y. Ng, and W. Pedrycz, "Ensembling perturbation-based oversamplers for imbalanced datasets," Neurocomputing, vol. 479, pp. 1–11, Mar. 2022, doi: 10.1016/J.NEUCOM.2022.01.049.

[36] Y. Bai, M. Tan, H. Cao, J. Tang, and Z. Liang, "Intelligent classification of carbonate reservoir quality using multisource geophysical logging and seismic data," IEEE Trans. Geosci. Remote Sens., 2022, doi: 10.1109/TGRS.2022.3140790.

[37] P. Ma, A. Mondal, B. A. Konomi, J. Hobbs, J. J. Song, and E. L. Kang, "Computer Model Emulation with High-Dimensional Functional Output in Large-Scale Observing System Uncertainty Experiments," https://doi.org/10.1080/00401706.2021.1895890, vol. 64, no. 1, pp. 65–79, 2021, doi: 10.1080/00401706.2021.1895890.

[38] S. Li, K. Zhou, L. Zhao, Q. Xu, and J. Liu, "An improved lithology identification approach based on representation enhancement by logging feature decomposition, selection and transformation," J. Pet. Sci. Eng., vol. 209, p. 109842, Feb. 2022, doi: 10.1016/J.PETROL.2021.109842.

[39] D. A. Wood, "Enhancing lithofacies machine learning predictions with gamma-ray attributes for boreholes with limited diversity of recorded well logs," Artif. Intell. Geosci., vol. 2, pp. 148–164, Dec. 2021, doi: 10.1016/J.AIIG.2022.02.007.

[40] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, B. N. Tackie-Otoo, and M. Y. Taki,

"Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions," J. Pet. Sci. Eng., vol. 208, p. 109244, Jan. 2022, doi: 10.1016/J.PETROL.2021.109244.

[41] Y. Zhang, X. Xu, Z. Li, R. Yi, C. Xu, and W. Luo, "Modelling soil thickness using environmental attributes in karst watersheds," CATENA, vol. 212, p. 106053, May 2022, doi: 10.1016/J.CATENA.2022.106053.

[42] M. Matinkia, R. Hashami, M. Mehrad, M. R. Hajsaeedi, and A. Velyati, "Prediction of permeability from well logs using a new hybrid machine learning algorithm," Petroleum, Mar. 2022, doi: 10.1016/J.PETLM.2022.03.003.

[43] Y. Bo, Q. Liu, X. Huang, and Y. Pan, "Real-time hard-rock tunnel prediction model for rock mass classification using CatBoost integrated with Sequential Model-Based Optimization," Tunn. Undergr. Sp. Technol., vol. 124, p. 104448, Jun. 2022, doi: 10.1016/J.TUST.2022.104448.

[44] D. ; Li, J. ; Zhao, Z. A. Liu, D. Li, J. Zhao, and Z. Liu, "A Novel Method of Multitype Hybrid Rock Lithology Classification Based on Convolutional Neural Networks," Sensors 2022, Vol. 22, Page 1574, vol. 22, no. 4, p. 1574, Feb. 2022, doi: 10.3390/S22041574.

[45] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220–239, May 2017, doi: 10.1016/J.ESWA.2016.12.035.

[46] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," IEEE Trans. Syst. Man, Cybern. Part ASystems Humans, vol. 40, no. 1, pp. 185–197, Jan. 2010, doi: 10.1109/TSMCA.2009.2029559.

[47] E. Elyan, C. F. Moreno-Garcia, and C. Jayne, "CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification," Neural Comput. Appl., vol. 33, no. 7, pp. 2839–2851, Apr. 2021, doi: 10.1007/S00521-020-05130-Z/TABLES/6.

[48] G. Kaur, “A comparison of two hybrid ensemble techniques for network anomaly detection in spark distributed environment,” J. Inf. Secur. Appl., vol. 55, p. 102601, Dec. 2020, doi: 10.1016/J.JISA.2020.102601.

[49] M. Luo et al., “Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II,” Energy Build., vol. 210, p. 109776, Mar. 2020, doi: 10.1016/J.ENBUILD.2020.109776.

[50] E. Florian, F. Sgarbossa, and I. Zennaro, “Machine learning-based predictive maintenance: A cost-oriented model for implementation,” Int. J. Prod. Econ., vol. 236, p. 108114, Jun. 2021, doi: 10.1016/J.IJPE.2021.108114.

[51] Z. Wu, W. Lin, and Y. Ji, “An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics,” IEEE Access, vol. 6, pp. 8394–8402, Feb. 2018, doi: 10.1109/ACCESS.2018.2807121.

[52] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, “Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning,” Int. J. Min. Sci. Technol., Sep. 2021, doi: 10.1016/J.IJMST.2021.08.004.

[53] K. Ostad-Ali-Askari and M. Shayannejad, “Computation of subsurface drain spacing in the unsteady conditions using Artificial Neural Networks (ANN),” Appl. Water Sci., vol. 11, no. 2, pp. 1–9, Feb. 2021, doi: 10.1007/S13201-020-01356-3/TABLES/4.

[54] A. K. Lysdahl et al., “Integrated bedrock model combining airborne geophysics and sparse drillings based on an artificial neural network,” Eng. Geol., vol. 297, p. 106484, Feb. 2022, doi: 10.1016/J.ENGGEO.2021.106484.

[55] A. Yadav, K. Yadav, and Anirbid Sircar, “Feedforward Neural Network for joint inversion of geophysical data to identify geothermal sweet spots in Gandhar, Gujarat, India,” Energy Geosci., vol. 2, no. 3, pp. 189–200, Jul. 2021, doi: 10.1016/J.ENGEOS.2021.01.001.

[56] S. Li, J. Chen, C. Liu, and Y. Wang, “Mineral Prospectivity Prediction via

Convolutional Neural Networks Based on Geological Big Data," J. Earth Sci. 2021 322, vol. 32, no. 2, pp. 327–347, Apr. 2021, doi: 10.1007/S12583-020-1365-Z.

[57] K. Dimililer, H. Dindar, and F. Al-Turjman, "Deep learning, machine learning and internet of things in geophysical engineering applications: An overview," Microprocess. Microsyst., vol. 80, p. 103613, Feb. 2021, doi: 10.1016/J.MICPRO.2020.103613.

[58] V. Harid et al., "Automated Large-Scale Extraction of Whistlers Using Mask-Scoring Regional Convolutional Neural Network," Geophys. Res. Lett., vol. 48, no. 15, p. e2021GL093819, Aug. 2021, doi: 10.1029/2021GL093819.

[59] B. Sadhukhan, S. Chakraborty, and S. Mukherjee, "Investigating the relationship between earthquake occurrences and climate change using RNN-based deep learning approach," Arab. J. Geosci. 2021 151, vol. 15, no. 1, pp. 1–27, Dec. 2021, doi: 10.1007/S12517-021-09229-Y.

[60] N. V. Korneev, J. V. Korneeva, S. P. Yurkevichyus, and G. I. Bakhturin, "An Approach to Risk Assessment and Threat Prediction for Complex Object Security Based on a Predicative Self-Configuring Neural System," Symmetry 2022, Vol. 14, Page 102, vol. 14, no. 1, p. 102, Jan. 2022, doi: 10.3390/SYM14010102.

[61] S. Raj, S. Tripathi, and K. C. Tripathi, "ArDHO-deep RNN: autoregressive deer hunting optimization based deep recurrent neural network in investigating atmospheric and oceanic parameters," Multimed. Tools Appl. 2022 816, vol. 81, no. 6, pp. 7561–7588, Jan. 2022, doi: 10.1007/S11042-021-11794-Z.

[62] W. De Mulder, S. Bethard, and M. F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," Comput. Speech Lang., vol. 30, no. 1, pp. 61–98, Mar. 2015, doi: 10.1016/J.CSL.2014.09.005.

[63] S.-B. Zhang, H.-J. Si, X.-M. Wu, and S.-S. Yan, "A comparison of deep learning methods for seismic impedance inversion," Pet. Sci., Jan. 2022, doi: 10.1016/J.PETSCI.2022.01.013.

[64] F. Aden-Antoniów, W. B. Frank, and L. Seydoux, "An Adaptable Random Forest

Model for the Declustering of Earthquake Catalogs," J. Geophys. Res. Solid Earth, vol. 127, no. 2, p. e2021JB023254, Feb. 2022, doi: 10.1029/2021JB023254.

[65] S. Sarkar and J. Maiti, "Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis," Saf. Sci., vol. 131, p. 104900, Nov. 2020, doi: 10.1016/J.SSCI.2020.104900.

[66] S. Bera, A. Das, and T. Mazumder, "Evaluation of machine learning, information theory and multi-criteria decision analysis methods for flood susceptibility mapping under varying spatial scale of analyses," Remote Sens. Appl. Soc. Environ., vol. 25, p. 100686, Jan. 2022, doi: 10.1016/J.RSASE.2021.100686.