iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76127

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

An Efficient Feature Selection Approach for Clinical Text Classification Using SFS and CatBoost

R. Gowthami[1], Dr. Ch. D.V. Subba Rao[2]
[1]M. Tech Student, [2]Professor, Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Tirupati, A.P

Abstract: ***Text classification in clinical settings is a very important and serious task in the sphere of Natural Language Processing, which has far-reaching implications in healthcare processes. We shall be categorising the medical transcripts into their respective medical conditions in this NLP project.***
To accomplish this, we employ the Sequential Forward Selection (SFS) method, a feature selection technique chosen specifically for its dimensional reduction capabilities. Using SFS, not only do we hope to increase the classification performance, but also to increase the efficiency of pattern recognition so that the disease can be detected not only fast, but also accurately. The contribution to the research highlights the importance of Clinical Text Classification, and in particular, how to streamline the process with the help of SFS.
Keywords: Clinical Text Classification, Natural Language Processing (NLP), Sequential Forward Selection (SFS), Feature Selection, Dimensionality Reduction, Medical Transcripts, Machine Learning.

I. INTRODUCTION

The classification of clinical texts is an urgent project in the field of healthcare necessitated by the desire to utilise the abundance of information held by unstructured clinical text materials. The issue has become prominent in this field because it is a multifaceted area with the potential to influence healthcare management and patient outcomes. We have suggested a procedure in this NLP project that is capable of categorising medical transcripts into the appropriate medical speciality. The medical speciality is the target variable, and the text information in the medical transcripts is the feature. There will be a number of critical steps in the project. Pre-processing of the text data will be required, which may include things like tokenisation, stemming and the removal of stop words. After pre-processing the data, there are several machine learning algorithms that are trained and tested on the data, including Logistic regression, Support Vector Machines, or Categorical Boosting. Accuracy, precision, recall, and F1-score are some of the metrics used to assess the performance of any model. Lastly, the most successful model can be chosen and implemented to categorise the new medical transcripts into their medical category.

II. PROBLEM DEFINITION

The Domain Need (Clinical Text Classification): The project addresses Clinical Text Classification, which is defined as an urgent task in Natural Language Processing (NLP), and there are important implications in healthcare applications. The motivating force behind this requirement is the wish to tap into the richness of information that exists in the form of unstructured clinical text information.

III. OBJECTIVES

The project aims to create an efficient and correct model in classifying the clinical text data by using Sequential Forward Selection (SFS) to select the features. The idea is to derive the most considered features of high-dimensional clinical data, and thus, obtain more interpretable models that require less computational resources and better classification. The project will assist in healthcare decision-making by helping extract insights from the unstructured clinical notes (medical notes, discharge notes, or diagnostic reports) in a faster and more reliable way.

IV. SCOPE

Categorising big and unstructured clinical text data to automate or support clinical decision-making, enhance patient outcomes, or simplify medical procedures. Sequential Forward Selection (SFS) shall be used as a procedure of choosing the most germane features of the text information, which is essential in lowering the dimensions as well as maximization of the model performance. Iteratively select the most useful features with the help of SFS to classify data. Compare the performance with the baseline methods (e.g., all features, or other feature selection techniques such as mutual information or chi-square).

V. RELATED WORK

Clinical Text Classification via Deep Learning and Rule-Based Features: Liang Yao, et al [1] deal with clinical text classification by stating that previous research in the field traditionally relied on rules or sources of knowledge-based feature engineering, and limited studies have utilised the effective feature learning properties of deep learning approaches. They offered a new method with rule-based properties and a knowledge-driven deep learning methodology to perform disease classification effectively. This technique was tested on the 2008 Integrating Informatics with Biology and the Bedside (i2b2) challenge on obesity, and the outcome was that the results of their type were found to be more efficient than the state-of-the-art approaches [1]

Deep Learning in Health Informatics: Ravi et al [2] talked about an application of data analytics in health informatics that has developed at an extremely high pace following an enormous volume of multimodality data. Artificial neural networks form the basis of deep learning, which is an emerging, powerful tool that is destined to remake the future of artificial intelligence. The rapid advances in computational power, high data storage speeds, and parallelisation have led to the rapid adoption of deep learning, as well as the fact that it automatically produces high-level features and semantic meaning out of input data, optimised automatically. The key areas that this extensive review concentrated on are the application of deep learning in translational bioinformatics, medical imaging, pervasive sensing, medical informatics, and public health [2].

Feature Selection Using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network: A. Marcano-Cedeño et al, [3] investigated the aspect of feature selection as an important method of data dimensionality reduction. Reduction of data is essential as it may enhance the performance of the classification, approximation functions, and the pattern recognition systems in speed, accuracy, and simplicity. They introduced a feature-selection process that is grounded on the Sequential Forward Selection (SFS) and a Feed Forward Neural Network (FFNN) that predicts the error as a selection criterion. The results of testing SFS-FFNN on Perceptron Multilayer (AMMLP) demonstrated that the achieved accuracy of classification was higher than that produced by the traditional Backpropagation (BP) algorithm and other recent feature selection algorithms applied on the same database. This is why the offered SFS-FFNN with AMMLP can be considered an interesting option to decrease the data dimensionality and obtain a high level of accuracy [3].

Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management, et al [4] examined the semi-supervised methods of learning to make use of the extensive volumes of unlabeled data within Electronic Medical Records (EMRs). They observed that the size of labelled training data is a challenge to developing machine-learning-based clinical text classifiers, because of the labour and cost involved in manually reviewing the data by clinicians. By comparison, EMR systems harbour hundreds of thousands of unlabeled notes that are typically not used by the supervised methods. They have compared the linear and Laplacian Support Vector Machines (SVMs) in a clinical text classification task. The Laplacian SVM, which used almost 20,000 randomly sampled non-labelled notes on top of the labelled training reference standard, yielded a much better performance than supervised SVMs. The quality of the performance increased as the number of labelled and non-labelled notes to train the Laplacian SVM increased. These findings indicate that semi-supervised approaches such as Laplacian SVM can be useful to utilise large and unlabeled corpora in EMRs to enhance clinical text classification [4].

What can natural language processing do for clinical decision support? Dina Demner-Fushman, et al [5] examined the application of NLP in Computerised Clinical Decision Support (CDS). CDS intends to support healthcare services and the general population by providing conveniently available health information whenever and wherever required. NLP plays an important role in the use of free-text information to power CDS, as a representation of clinical knowledge, and the ability to use clinical narrative in standardised formats.

The review was oriented on the recent revived interest in the development of the basic approaches toward NLP and the progress of NLP systems aimed at CDS development, taking into consideration the issues related to specific sublanguages, target audience, and support objectives [5].

VI. ALGORITHMS AND METHODOLOGY

The given system suggests the use of machine learning to automate the process of medical transcription classification using the latest NLP methods, the CatBoostClassifier, and the Sequential Forward Selection (SFS) to select features.

A. Sequential Forward Selection (SFS)

SFS is an algorithm of feature selection which is made to selects the optimal subset of features in a machine learning model by the use of repetition. It is especially applicable to optimising model performance by dimensionality reduction of data, which is useful in computational efficiency, and overfitting is alleviated, particularly with high-dimensional text data. SFS uses a greedy search technique. The initial step is a set of features that is empty. The algorithm measures the performance of the models at every step by including one feature into the existing pool. The best improvement in performance (according to a measure such as accuracy or F1 score) is included in the chosen set. The process persists until reaching a defined endpoint, such as a target feature count or when performance enhancement becomes negligible. SFS is a wrapper algorithm, that is, it uses a predictive model to evaluate the utility of subsets of features.

B. CatBoost Classifier

CatBoost (Categorical Boosting) is a machine learning algorithm written in open source by Yandex, which is based on gradient boosting. It is particularly created to support categorical features without the need to perform a lot of pre-processing, such as one-hot encoding. CatBoost does this by an advanced encoding strategy that makes use of count-based encoding and target statistics to produce effective numerical encodings of categorical variables. It is sequential in its decision tree building to reduce the residual error. More importantly, CatBoost relies on a process known as ordered boosting. The approach splits the information into segments to avoid data contamination, creating more robust models that remain effective even with limited sample sizes. The main characteristics of CatBoost are its high performance, robustness, effectiveness of default parameters, and native support of missing data. The implementation was done with the following model parameters: iterations=200, learningrate=0.1, and depth=6.

VII. METHODOLOGY

In this section, we describe the system architecture that is proposed and the approach that we adopted in this study.

A. System Architecture

The suggested architecture diagram of the system is in Figure 1. Presents the System Architecture of this project clearly and conceptually captured in the system methodology and module definitions.

B. System Architecture Overview

The architecture will handle the unstructured clinical text input data feeds through the preprocessing and feature selection steps in an organised manner, and finally, the machine learning.

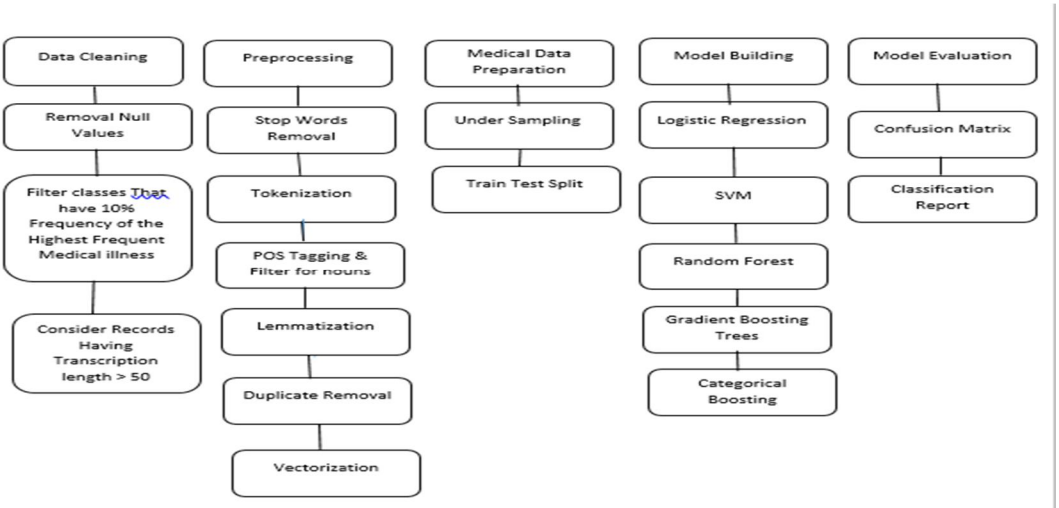


Fig.1. SYSTEM ARCHITECTURE

The learning model will classify the input data automatically. The system is based on the web application framework (suggested by the use of HTML/CSS as the front-end and Django in the code) that handles user authentication, data viewing, training, and prediction.

The design of the project includes the following parts and steps:

1) *Authentication Layers and System Design:* The system employs a multi-layered system that involves users and administrators.
2) *User Registration and Login:* The user needs to first register and then log in with valid credentials.
3) *Admin Activation:* The Admin is obliged to log in and activate the registered user account; until the status of the user account is activated, the user will not be able to use the features.
4) *Input Design:* The input design is concerned with security, ease, prevention of errors, and maintenance of privacy. It directs the working staff on how to give input, such as the validation means.
5) *Output Design:* The output design dictates the manner in which the processed information (such as the prediction result) is presented transparently to the user so as to enhance decision-making.

VIII. IMPLEMENTATION DETAILS

The main functional architecture is based on the Clinical Data Analysis Application method.

1) *Dataset Collection (Input):* The system obtains the data (a mtsamples.csv file) that was obtained via such sources as the Kaggle website.
2) *Pre-processing:* This module takes care of the unstructured raw clinical text data. It detects and replaces missing values (NaN) using techniques like ffill (forward fill). Preprocessing of text involves lemmatisation, lowercasing, stop-word removal and tokenisation. All textual elements (description, transcription, samplename, keywords) are consolidated into a single feature: combinedtext**.** Vectorisation: TF-IDF is used to transform the combined text data into numerical forms (that is, it transforms the text into numbers), commonly with a specified maximum number of features (e.g., max out features=1000).
3) *Feature Selection (SFS):* This is the main architectural innovation in order to deal with the high dimensionality of clinical text. The architecture takes advantage of Sequential Forward Selection (SFS), a wrapper-based method. SFS selects the most relevant features iteratively depending on the contribution they make to the model. It is aimed at simplifying the dimensions of data, enhancing the performance of classification, decreasing the complexity of computations and increasing the interpretability.
4) *Training and Model Selection:* The data in the form of vectors is divided into the training and the testing data (e.g., 80% and 20%). On the training set, RANDOM UnderSampler (RUS) is applied to balance the training set, followed by model fitting. The first classification algorithm is the CatBoostClassifier. The model is trained with the help of the fit method. The measures applied in model evaluation are accuracy, precision, recall, and F1-score. The model/vectorizer is preserved (joblib. dump) for deployment purposes.
5) *Prediction and Deployment:* Previously trained model and vectorizer are retrieved from saved files (.pkl).

Description, transcription, sample name, and keywords are input by the user. The stored vectorizer processes and transforms the input into vectors. The CatBoost model is used to predict medical specialities. The model is implemented in a Clinical web application.

IX. EVALUATION AND RESULTS

The trained CatBoost model was tested on the resampled training dataset to check how well it learned. It reached a training accuracy of 99.15%.

Fig 9.1 The User can view the home page.

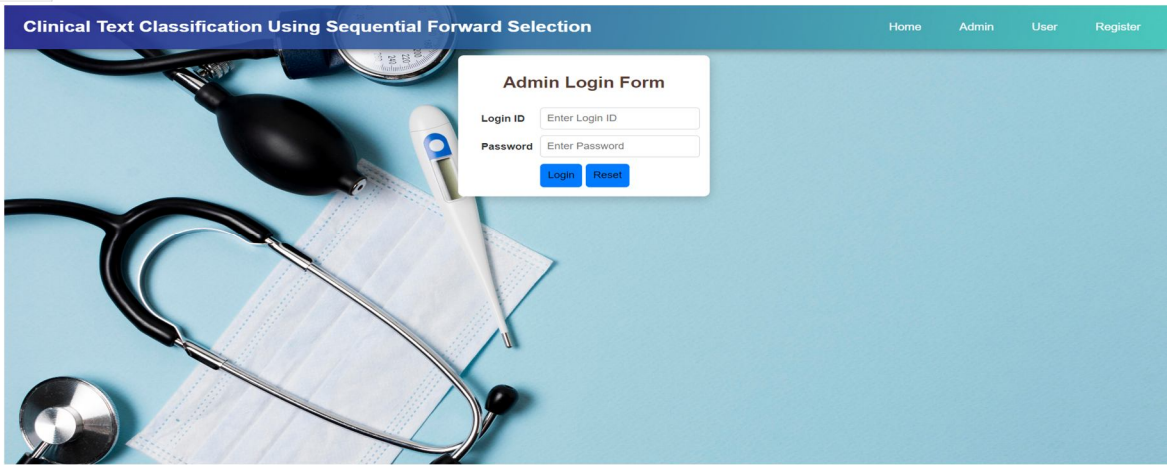


Fig 9.2 Admin login page

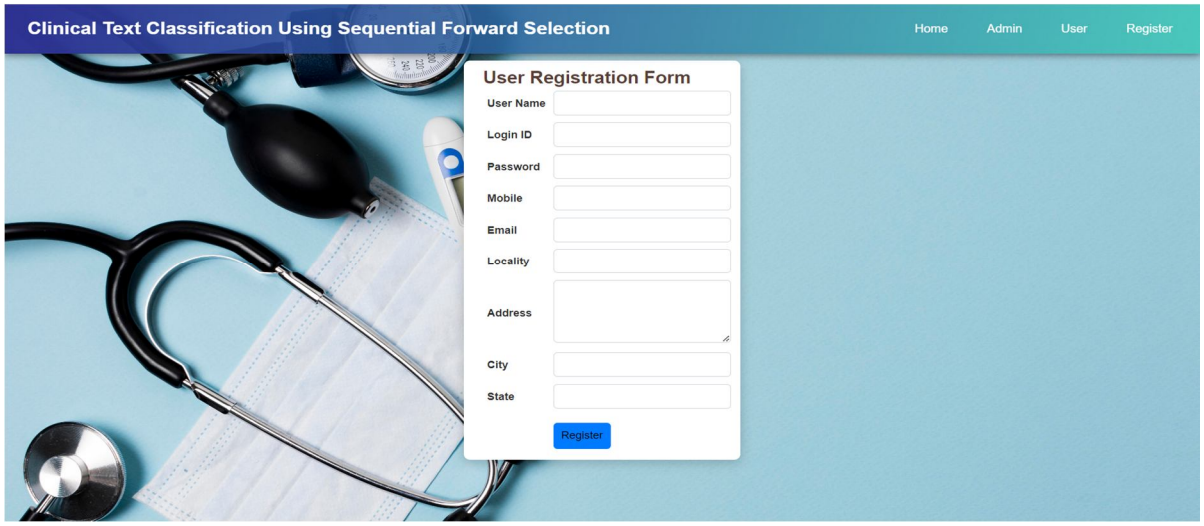


Fig 9.3 User can register with required details.

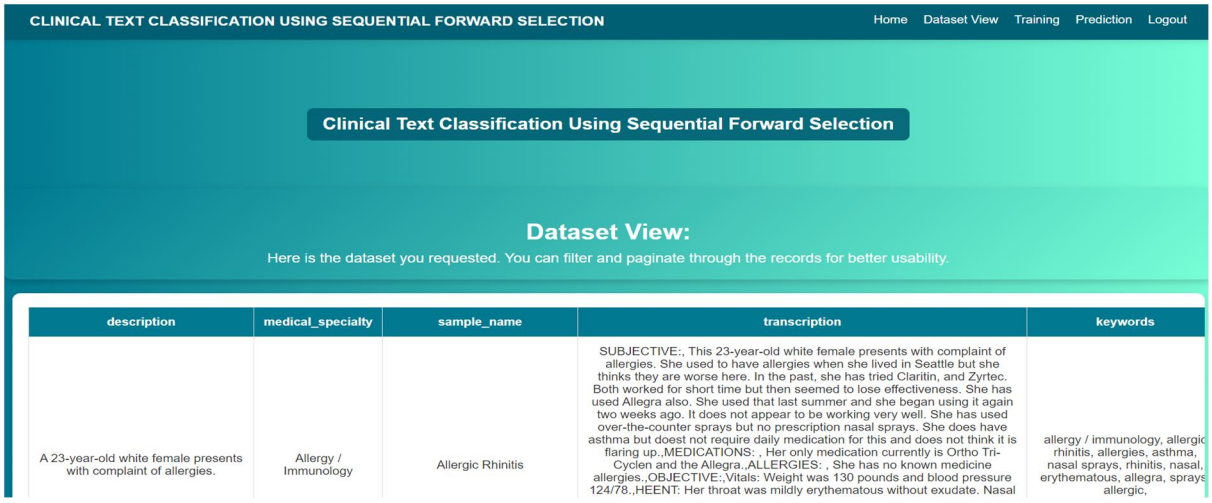


Fig 9.4. The user can view the data

Fig 9.5 Training Dataset

Fig 9.6 The User can give an input and view The Predicted Result.

X. CONCLUSION

This study shows that sub-categorizing similar groups within the dataset can simplify the classification process by reducing the number of categories to be analyzed. Although manually engineered features may improve performance on this dataset, they are unlikely to generalize well to other clinical transcription datasets. Our findings indicate that additional data is necessary to classify the transcriptions accurately into their respective medical categories, as the limited size of the current dataset restricts the achievable accuracy.

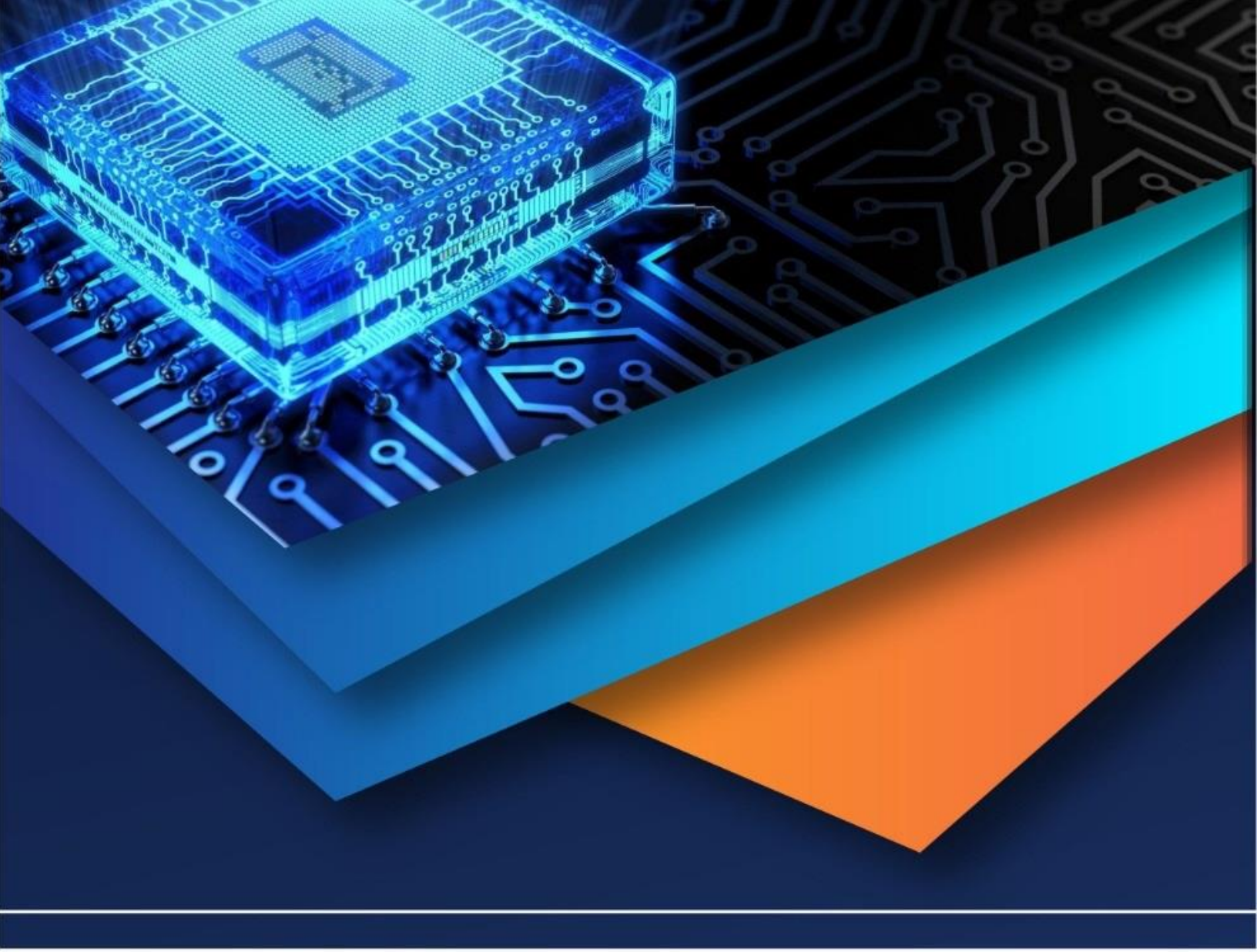
XI. SCOPE FOR FUTURE WORK

Future work will focus on expanding the dataset and dividing the transcriptions into smaller, more specific units. This will support the use of multi-class or multi-label classification methods, enabling more detailed and precise categorization of medical content. These enhancements are expected to significantly improve the overall quality and effectiveness of the medical transcription classification system.

REFERENCES

[1] Yao, L., Mao, C., and Luo, Y. developed a method combining rule-based features with knowledge-driven CNNs to classify clinical text. This study was published in BMC Medical Informatics and Decision Making, in the 19th volume, Supplement 3, on page 71, in 2019.
[2] Ravi D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G. Z. explored the use of deep learning methods in health informatics. Their research appeared in the IEEE Journal of Biomedical and Health Informatics, Volume 21, Issue 1, spanning pages 4–21, in 2017.
[3] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina Januchs, and D. Andina wrote about feature selection by using Sequential Forward Selection and a classification approach with Artificial Metaplasticity Neural Network. Their work was presented at the IECON 2010 - 36th Annual IEEE Industrial Electronics Society Conference held in Glendale, Arizona, USA, pages 2845-2850 in 2010.
[4] V. Garla, C. Taylor, and C. Brandt discussed semi-supervised classification of clinical text through Laplacian SVMs. They showed its use in managing cancer cases. Their findings appeared in the Journal of Biomedical Informatics, volume 46, issue 5, pages 869–875, published in 2013.
[5] Demner-Fushman D., Chapman, W. W., and McDonald, C. J. How does natural language processing help in clinical decision-making? Journal of Biomedical Informatics 42(5) 760–772, 2009.

[6] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Finding patient smoking habits using medical discharge data. Journal of the American Medical Informatics Association 15(1) 14–24, January 2008.
[7] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. They explored Medical Semantic Similarity using a Neural Language Model. This study was presented at the 23rd ACM International Conference on Information and Knowledge Management (CIKM '14). The Association for Computing Machinery published it in New York, NY, USA, on pages 1819 to 1822, in 2014.
[8] M. Last, A. Kandel, and O. Maimon worked on an information-theoretic algorithm to select features. Their research appears in the journal Pattern Recognition Letters Volume 22, Issues 6-7, on pages 799 to 811, published in 2001.
[9] Kudo, Mineichi and Sklansky, Jack. J. Sklansky. Comparison of Methods to Pick Features in Pattern Classifiers. Pattern Recognition Volume 33, Pages 25–41, Year 2000.
[10] D. Xiao and J. Zhang. The Role of Feature Importance and Choosing Features. Published in the proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY