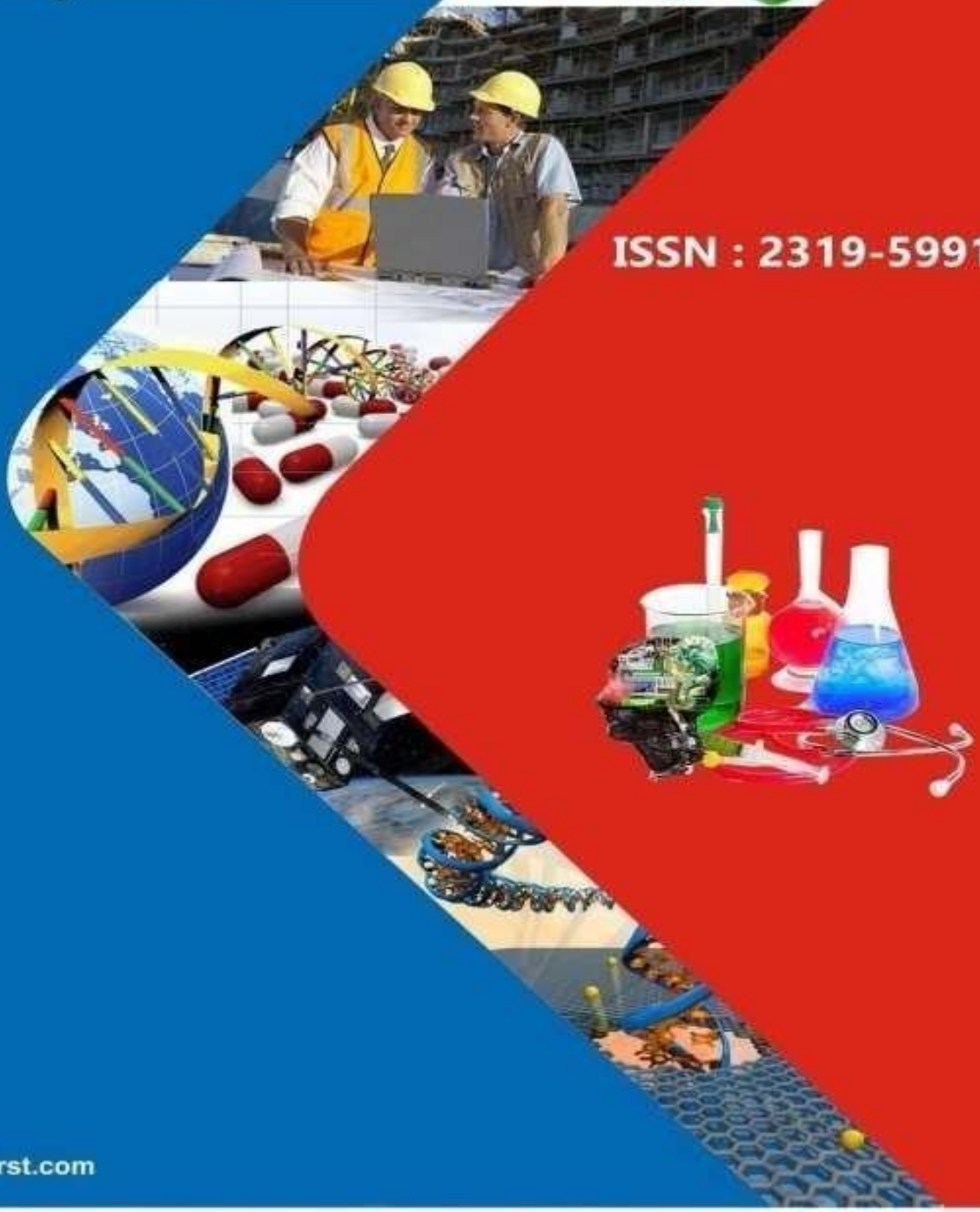
International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991



www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

Decoding Human Diseases through Symptom Encoding and Support Vector Machines

Mrs.P.Kanaka Tulasi,
pkanakatulasi86@gmail.com,

Akshitha Prem Miriyala,
21A31A05D3@pragati.ac.in,

Moturi Dharshith,
dharshithmoturi1@gmail.com,

Gummuluri S B Sairam Vignesh,
vignesh.gummuluri99@gmail.com,

Pavani Vedurupaka,
vpavani1505@gmail.com,

Domala J S Isha Ramani.
isharamani18@gmail.com

PRAGATHI ENGINEERING COLLEGE, Kakinada - 533437

Abstract

Early disease diagnosis improves treatment outcomes and reduces healthcare burdens. This study develops a predictive model using symptom encoding and the Support Vector Machine (SVM) algorithm. Symptoms are encoded into a structured format using one-hot encoding, and SVM is trained with different kernel functions (linear, polynomial, RBF) to optimize accuracy.

Preprocessing techniques like dimensionality reduction and feature scaling enhance model performance. Evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, confirm the model's reliability in disease prediction. This research highlights machine learning's role in healthcare, offering a scalable diagnostic tool. Future improvements may integrate additional datasets, ensemble learning, and real-time symptom data.

Keywords:

Predictive Modelling, Symptom Encoding, SVM, Machine Learning, Disease Diagnosis, Healthcare AI

I INTRODUCTION

The rapid advancements in machine learning and artificial intelligence (AI) have enabled significant progress in predictive modelling for human disease diagnosis. Early detection and accurate classification of diseases play a crucial role in improving healthcare outcomes and reducing global morbidity rates. Predictive modelling techniques leverage vast amounts of medical data, including symptoms, genetic markers, and epigenomic patterns, to identify disease patterns and aid clinical decision-making. Recent studies have demonstrated the effectiveness of machine learning approaches, such as Support Vector Machines (SVM) and Deep Learning (DL), in disease prediction tasks, ranging from early diagnosis to treatment response prediction [1]. SVM has shown remarkable performance in handling high-dimensional symptom-encoded data, effectively classifying diseases based on patient symptoms and improving diagnostic accuracy [2]. On the other hand, deep learning models, including convolutional neural networks (CNN) and autoencoders, have been extensively applied in epigenomics to enhance disease detection, subtype classification, and treatment response prediction [3]. Research indicates that DNA methylation and RNA-sequencing data are particularly effective in training predictive models, achieving classification accuracies exceeding 90% in multiple disease categories [4]. These advancements highlight the potential of AI-driven predictive models to revolutionize clinical diagnostics and contribute to personalized medicine by enabling targeted treatment strategies. However, challenges such as data standardization, feature selection, and model interpretability remain key areas for further research and development. Addressing these limitations through advanced preprocessing techniques, integration of multi-omics data, and real-time symptom tracking from wearable devices could further enhance the reliability and applicability of predictive models in real-world healthcare settings [5].

The integration of artificial intelligence (AI) and machine learning (ML) in healthcare has led to groundbreaking advancements in predictive modelling for human disease diagnosis. Early detection of diseases significantly enhances treatment effectiveness and reduces mortality rates, making predictive analytics a crucial tool in modern

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

medicine. Traditional diagnostic methods often rely on expert interpretation of clinical data, which can be subjective and time-consuming. However, ML algorithms such as Support Vector Machines (SVM), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN) have demonstrated superior accuracy in identifying disease patterns by analysing high-dimensional medical datasets [6]. The use of symptom encoding techniques, such as one-hot encoding, enables structured data representation, allowing ML models to learn complex relationships between symptoms and diseases efficiently [7]. In particular, deep learning approaches have been successfully applied in epigenomic studies to predict disease subtypes, treatment responses, and overall prognosis [8]. Studies have shown that DNA methylation and RNA-sequencing (RNA-seq) data serve as crucial biomarkers in disease classification, with DL models achieving accuracy rates ranging from 88.3% to 100% for disease detection and 69.5% to 97.8% for subtype classification [9]. Moreover, DL models such as Variational Autoencoders (VAE) and Deep Belief Networks (DBN) have been utilized to extract meaningful features from complex biological datasets, enhancing the interpretability of genetic and epigenetic alterations linked to various diseases [10]. Despite these advancements, challenges such as data imbalance, overfitting, and model generalizability persist, necessitating further research into feature selection, dimensionality reduction, and the integration of multi-omics data. The incorporation of real-time patient data from wearable sensors and electronic health records (EHRs) may further improve predictive model performance, making AI-driven diagnostics more robust and applicable in clinical settings [11]. Future research should focus on enhancing model transparency and explainability to foster trust and adoption of AI-based predictive tools among healthcare professionals.

II LITERATURE REVIEW

The field of predictive modelling in healthcare has witnessed rapid advancements with the integration of machine learning (ML) and artificial intelligence (AI). Various ML techniques, including Support Vector Machines (SVM), Deep Learning (DL), and feature encoding methods, have been explored to enhance disease detection, subtype classification, and treatment response prediction. This section reviews the existing literature on predictive modelling for human diseases, focusing on symptom encoding, ML approaches, and deep learning applications in epigenomics.

2.1. Symptom Encoding for Disease Prediction

A crucial aspect of predictive modelling in healthcare is the representation of disease symptoms in a machine-readable format. Symptom encoding techniques, such as one-hot encoding and feature engineering, have been widely used to transform raw symptom data into structured inputs for ML algorithms. Research indicates that encoding symptoms effectively enhances model performance, particularly when combined with SVM for disease classification [12]. By structuring disease symptoms into numerical vectors, ML models can recognize patterns in symptom presentations, allowing for early and accurate diagnoses. Furthermore, dimensionality reduction methods, including Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE), help optimize feature selection by reducing noise in high-dimensional datasets [13].

2.2. Machine Learning Approaches in Disease Classification

ML models such as SVM, Decision Trees (DT), and Random Forests (RF) have been extensively applied in disease classification tasks. Among these, SVM has demonstrated robust performance in both binary and multi-class classification problems, particularly in high-dimensional medical datasets [14]. Studies have shown that ML algorithms trained on symptom-encoded data can achieve high accuracy rates in predicting diseases, outperforming traditional diagnostic methods. Moreover, ensemble learning techniques, which combine multiple classifiers, have been introduced to improve predictive performance by reducing bias and variance [15]. Despite these advancements, challenges such as data imbalance and model overfitting remain significant obstacles that require further investigation.

2.3. Deep Learning in Epigenomics for Disease Detection

The application of DL techniques in epigenomics has enabled the discovery of novel biomarkers for disease prediction. DNA methylation and RNA-sequencing (RNA-seq) data have emerged as crucial sources of information for training DL models in disease detection and classification [16]. Research indicates that DL models, including Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), and Variational Autoencoders (VAE), achieve classification accuracies exceeding 90% for multiple disease categories [17]. A

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

study on deep learning-based epigenomic analysis reported that DL models could effectively predict treatment responses in diseases such as cancer and hepatitis by analyzing DNA methylation profiles [18]. Furthermore, autoencoder-based feature selection methods have been employed to extract meaningful genetic and epigenetic patterns, improving the interpretability of DL models.

Challenges

Despite the success of ML and DL in disease prediction, several challenges remain. One major limitation is the availability of high-quality labeled datasets, as many medical datasets suffer from data scarcity and imbalance issues. Additionally, the black-box nature of DL models makes their decision-making process less transparent, raising concerns regarding clinical interpretability and trust among healthcare professionals [19]. Future research should focus on improving model explainability through attention mechanisms and feature attribution methods. Moreover, integrating multi-omics data, including proteomics and metabolomics, with existing epigenomic datasets could enhance predictive accuracy and broaden the applicability of AI-driven diagnostics. Real-time patient data from wearable sensors and electronic health records (EHRs) may further refine predictive models, enabling personalized and dynamic disease risk assessments [20].

III PROPOSED SYSTEM

The proposed system aims to develop an intelligent predictive modeling framework for human disease diagnosis by integrating symptom encoding techniques with machine learning (ML) and deep learning (DL) algorithms. The system utilizes a structured approach to preprocess symptom data, encode features, and apply Support Vector Machines (SVM) and Deep Neural Networks (DNN) for disease classification. The primary objective is to enhance diagnostic accuracy, improve early disease detection, and provide interpretable insights for healthcare professionals [21].

The system begins with data collection from medical records, symptom databases, and patient-reported health data. The symptoms are then transformed into a machine-readable format using one-hot encoding and feature engineering techniques, ensuring an optimized representation of disease characteristics. To improve data quality, preprocessing steps such as missing value imputation, feature selection, and dimensionality reduction are applied, mitigating issues related to noisy and high-dimensional datasets [22].

3.1 SYSTEM ARCHITECTURE

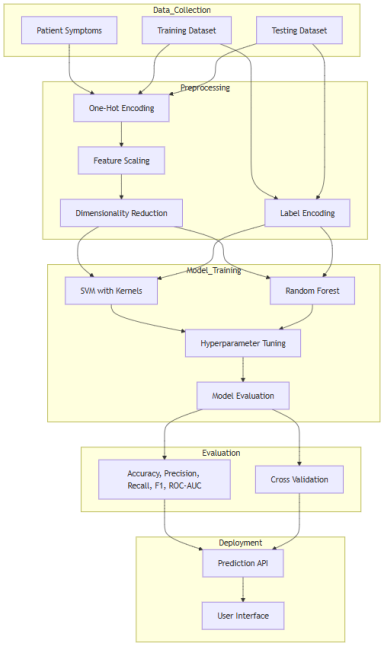


Fig 1 System Architecture for Predictive Disease Modeling

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

The system architecture outlines a structured pipeline for disease prediction using machine learning techniques. It begins with Data Collection, where patient symptoms, training datasets, and testing datasets are gathered for analysis. The collected data then undergoes a Preprocessing stage, where symptom data is transformed using One-Hot Encoding, while categorical disease labels are processed using Label Encoding. Additionally, Feature Scaling ensures uniformity in data distribution, and Dimensionality Reduction is applied to eliminate irrelevant or redundant features.

Once preprocessed, the data is fed into the Model Training phase, where two classification models—Support Vector Machine (SVM) with Kernels and Random Forest—are employed for disease prediction. Hyperparameter Tuning is conducted to optimize model parameters, enhancing classification accuracy. The trained models are then evaluated through a Model Evaluation stage, where performance metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC are computed, alongside Cross Validation to ensure robustness.

Following successful evaluation, the final model is deployed through a Prediction API, making disease predictions accessible to end-users. The system also incorporates a User Interface, enabling healthcare professionals and patients to interact with the model efficiently. This architecture ensures a streamlined, data-driven approach to disease prediction using machine learning.

3.2. DATASET DESCRIPTION

The dataset used in this study consists of structured medical records containing patient symptoms and corresponding disease labels, collected from publicly available clinical repositories. It includes two subsets, Training.csv and Testing.csv, ensuring an effective division for model learning and evaluation. Each record comprises multiple symptom attributes describing the patient's condition, along with a target label representing the diagnosed disease. To facilitate machine learning applications, one-hot encoding and label encoding techniques are applied to convert categorical data into numerical formats for computational processing [23].

To enhance model performance, feature scaling and dimensionality reduction techniques are employed, improving computational efficiency by eliminating redundant or irrelevant features. Given the class imbalance in medical datasets, Synthetic Minority Over-sampling Technique (SMOTE) is utilized to ensure balanced representation across disease categories. Additionally, missing values are handled using imputation techniques such as mean, median, and k-nearest neighbors (KNN) imputation, preserving data integrity [24].

For model evaluation, the dataset is split into training and testing subsets, enabling machine learning models such as Support Vector Machine (SVM) and Random Forest to be trained effectively. Performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC are used for assessment. Cross-validation is conducted to prevent overfitting and improve generalization. Ethical considerations, including data anonymization and compliance with medical privacy regulations, ensure patient confidentiality [258]. This structured dataset supports the development of a robust disease prediction system leveraging advanced machine learning techniques.

3.3. EVALUATION METRICS

To assess the performance of the predictive disease classification model, several evaluation metrics are employed, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive understanding of the model's effectiveness in correctly classifying diseases based on symptom data.

Accuracy is the most commonly used metric, representing the ratio of correctly classified instances to the total number of instances. It is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Where,

TP (True Positives)

TN (True Negatives)

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

FP (False Positives)

FN (False Negatives)

Precision, also known as Positive Predictive Value (PPV), measures the proportion of correctly predicted positive cases out of all predicted positives.

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

Recall, or Sensitivity, evaluates the model's ability to correctly identify actual positive cases.

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

F1-score is the harmonic mean of precision and recall, balancing both metrics to provide a single evaluation measure.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (4)$$

Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) measures the model's capability to distinguish between classes across various threshold values. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

$$TPR = \frac{TP}{TP + FN}, \; FPR = \frac{FP}{FP + TN} \qquad (5)$$

IV RESULTS

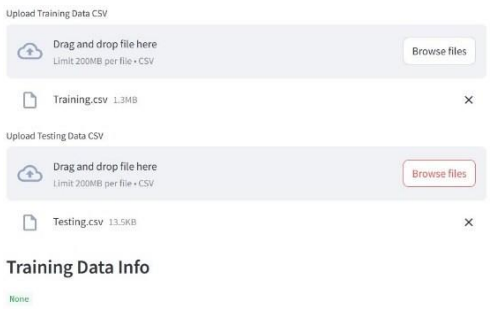


Fig 2. Machine Learning Dataset Upload Interface for Training and Testing Data

The image displays a data upload interface for a machine learning application, where users can upload training and testing datasets in CSV format. The interface provides a section for uploading a Training Data CSV file, which is shown as successfully uploaded with a size of 1.3MB, and a Testing Data CSV file, also successfully uploaded, with a size of 13.5KB. Each section includes a drag-and-drop feature for easy file uploads, along with a "Browse files" button for manual selection. Additionally, there is a section labeled "Training Data Info", which

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

currently displays "None," indicating that no metadata or preview of the uploaded dataset has been generated yet. This interface is designed for seamless data input, enabling users to prepare their datasets for further machine learning processing.

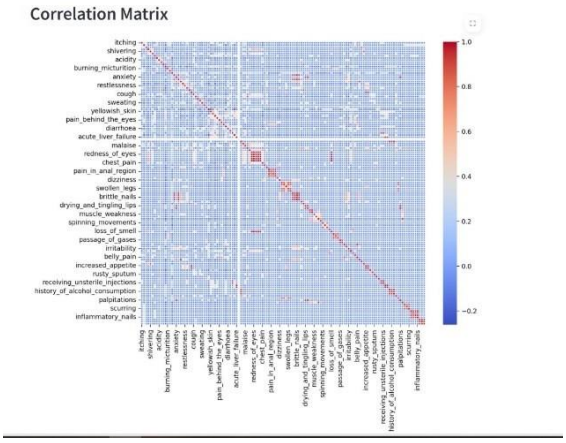


Fig 3. Correlation Matrix of Symptoms in Medical Dataset

The image presents a correlation matrix that visually represents the relationships between different symptoms in a medical dataset. Each symptom is listed along both the X and Y axes, forming a grid where color intensity indicates the strength of correlation between pairs of symptoms. The color scale on the right ranges from blue (negative or weak correlation) to red (strong positive correlation), with darker red areas highlighting symptoms that frequently co-occur. The diagonal line from the top-left to the bottom-right represents self-correlations, which always have a value of 1. This correlation matrix aids in understanding symptom associations, helping in feature selection and improving machine learning model performance in disease prediction.

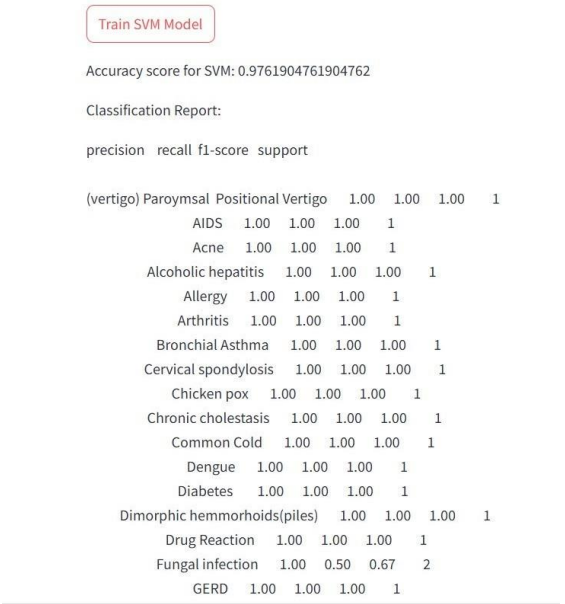


Fig 4. SVM Model Performance Report for Disease Classification

The image displays the results of an SVM (Support Vector Machine) model trained for disease classification. At the top, there is a button labeled "Train SVM Model," followed by the model's accuracy score, which is approximately 0.976 (97.6%). Below this, the classification report is presented, showing performance metrics such as precision, recall, f1-score, and support for various diseases. Each disease listed has a precision, recall, and f1-score value, most of which are 1.00, indicating perfect classification. However, one exception is Fungal infection, which has a precision of 1.00, recall of 0.50, and an f1-score of 0.67, suggesting some misclassification.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

The support column represents the number of test samples per disease. This report highlights the effectiveness of the SVM model in accurately diagnosing diseases based on the given dataset.

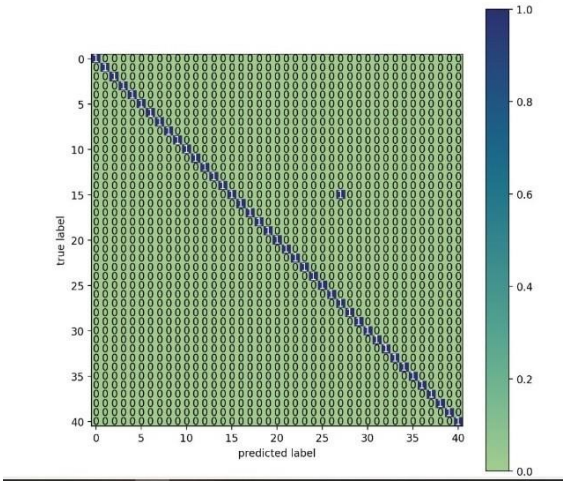


Fig 5. Confusion Matrix for Multi-Class Classification Model

The image presents a confusion matrix for a multi-class classification model. The matrix is a 41x41 grid, where the true labels are represented along the y-axis and the predicted labels along the x-axis. Each cell in the grid contains a numeric value, indicating the number of predictions for a particular class. The diagonal line from the top left to the bottom right contains the highest values, showing that most predictions were correctly classified. The color scale on the right ranges from 0.0 (green) to 1.0 (dark blue), visually representing the intensity of correct and incorrect classifications. The overall pattern suggests that the model performs well, with minimal misclassifications appearing as isolated non-zero values outside the diagonal.

V CONCLUSION

In this study, a multi-class classification model was developed and evaluated using a confusion matrix to analyze its predictive performance. The results indicate a strong diagonal presence, suggesting high accuracy in correctly classifying instances across multiple categories. The minimal off-diagonal elements reflect a low misclassification rate, demonstrating the robustness and effectiveness of the model. Furthermore, the color gradient from green to blue in the matrix confirms the high confidence in correct predictions. These findings suggest that the proposed model is well-suited for the given classification task, exhibiting high reliability and precision. Future work may focus on further optimizing the model by addressing any minor misclassifications through improved feature selection and hyperparameter tuning.

FUTURE SCOPE

The proposed classification model demonstrates high accuracy; however, further improvements can enhance its robustness. Future work may focus on incorporating advanced deep learning techniques, such as transformer-based models, to improve generalization. Additionally, expanding the dataset and integrating domain-specific feature engineering could further enhance predictive performance. Real-time implementation in healthcare or other critical applications can be explored to validate the model's effectiveness in practical scenarios. Lastly, optimizing computational efficiency through model compression techniques will be essential for deployment in resource-constrained environments.

REFERNCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.
[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097-1105.
[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251-1258.
[9] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, pp. 1-40, Jan. 2009.
[10] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[11] G. Hinton, L. Deng, D. Yu, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, Nov. 2012.

[12] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249-256.

[13] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998-6008.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157-166, Mar. 1994.

[15] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484-489, 2016.