A Course Project report submitted

in partial fulfilment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

Gurram Shravani - 2203A52024

Under the guidance of

Dr. D. RAMESH

Assistant Professor, School of CS&AI.

SR UNIVERSITY

SR University, Ananthsagar, Warangal, Telangana -506371

CONTENTS

CHAPTER 1

DATASET OVERVIEW

Project -1

The **Credit Card Fraud Detection dataset** on Kaggle contains credit card transactions made by European cardholders, collected over a two-day period in September 2013. It includes a total of **284,807 transactions**, out of which only **492 are fraudulent**, representing just **0.172%** of the dataset. Due to the extreme class imbalance, it serves as a challenging benchmark for fraud detection models.

To preserve privacy, the original features have been transformed using **Principal Component Analysis (PCA)**, and are labeled as **V1 to V28**. Only two features remain in their original form:

- **Time**: The number of seconds elapsed between each transaction and the first transaction in the dataset.
- **Amount**: The transaction amount, which can be useful for cost-sensitive learning.

The **'Class'** column is the target variable, with **1 indicating a fraudulent transaction** and **0 indicating a legitimate one**. This dataset is widely used in machine learning and anomaly detection projects, particularly for developing and evaluating algorithms that can detect fraud in highly imbalanced data environments.

.

Project – 2

The **Brain Tumor Classification (MRI)** dataset is designed for the classification of brain tumors using MRI (Magnetic Resonance Imaging) scans. It contains a total of **7,026 labeled images**, divided into four distinct categories of brain tumors:

- **Glioma Tumor**
- **Meningioma Tumor**
- **Pituitary Tumor**
- **No Tumor (Healthy Brain)**

The images are organized into training and testing sets, with each image provided in JPEG format. The dataset covers various MRI orientations and captures a wide range of visual features, making it suitable for training deep learning models such as CNNs (Convolutional Neural Networks) for image classification tasks.

This dataset is particularly useful for research and educational purposes in the fields of medical imaging, computer vision, and artificial intelligence. It can help develop automated diagnostic tools to assist radiologists and healthcare professionals in early detection and classification of brain tumours.

Project – 3

The **Real / Fake Job Posting Prediction t**his dataset contains job postings collected from various sources, designed to help detect **fake job listings**. With a total of **~18,000 job posts**, the dataset includes a mix of **real and fraudulent job advertisements**, labelled accordingly. It aims to support machine learning models in learning how to distinguish between legitimate and deceptive postings.

Each entry contains multiple fields such as:

- **Title**
- **Location**
- **Company**
- **Description**
- **Requirements**
- **Industry**
- **Function**
- **Employment Type**, and more.
-

The target variable is **'fraudulent'**, where:

- **0** = Real job posting
- **1** = Fake job posting

CHAPTER – 2

METHODOLOGY

Project – 1

Dataset Preparation:

The Credit Card Fraud Detection dataset includes 284,807 transactions, with only 492 labeled as fraud (0.172%), making it highly imbalanced. Features V1–V28 are anonymized using PCA, while 'Time' and 'Amount' remain original, and the target variable 'Class' indicates fraud or legitimate transactions.

Data Preprocessing:

Outlier removal was performed using the IQR method while ensuring that each class retained a minimum number of samples, preserving the integrity of the minority (fraud) class. This helped reduce noise without discarding important data points. After cleaning, the class imbalance was addressed using **SMOTE (Synthetic Minority Over-sampling Technique)**, which generated synthetic examples for the minority class. This preprocessing step ensured a more balanced and reliable dataset for model training.

Feature Selection:

For feature selection, all available columns were retained as relevant features for the model. Since the dataset's columns provide valuable information for detecting fraud, no features were dropped, ensuring the model leverages all available data for optimal performance.

Model Training:

Different machine learning models like Logistic Regression, Random Forest, and SVM were trained on the processed dataset. These models were evaluated based on accuracy, precision, recall, and F1-score.

Performance Evaluation:

Confusion matrices and metric scores were used to analyze how well each model detected network intrusions. Ensemble models such as Gradient Boosting showed balanced and effective results in identifying attacks.

Project -2

Dataset:

This dataset contains MRI images for brain tumor classification, including categories such as benign, malignant, and pituitary tumors. It is used to develop models that can assist in detecting brain tumors accurately, helping improve diagnostic efficiency and reduce human error.

Preprocessing:

Images resized to 256 x 256 pixels, normalized between 0-1, with augmentation techniques like rotation, flipping, and zooming applied.

Model Architecture:

CNN with convolutional, max-pooling, and dropout layers to extract features and reduce overfitting.

Training:

Categorical cross-entropy loss, with a separate validation set for performance monitoring.

Evaluation Metrics:

Accuracy, confusion matrix, and classification reports (precision, recall, F1-score) to assess model performance.

Project – 3

Dataset Preparation:

This dataset contains job postings labeled as real or fake, aimed at detecting fraudulent listings. It includes various fields like title, location, company, and description.

Preprocessing:

All text samples were cleaned by removing punctuation, converting to lowercase, removing stopwords, and tokenizing. Sequences were padded to ensure consistent input length.

Feature Extraction:

Text was converted into sequences using Keras Tokenizer and embedded into dense vectors to capture word semantics.

Model Architecture:

A deep LSTM model was developed to learn temporal patterns in the text. The model includes dropout layers to reduce overfitting and ends with a softmax layer for multiclass prediction.

Model Training:

The model was trained using categorical cross-entropy loss and Adam optimizer, with early stopping applied to prevent overfitting.

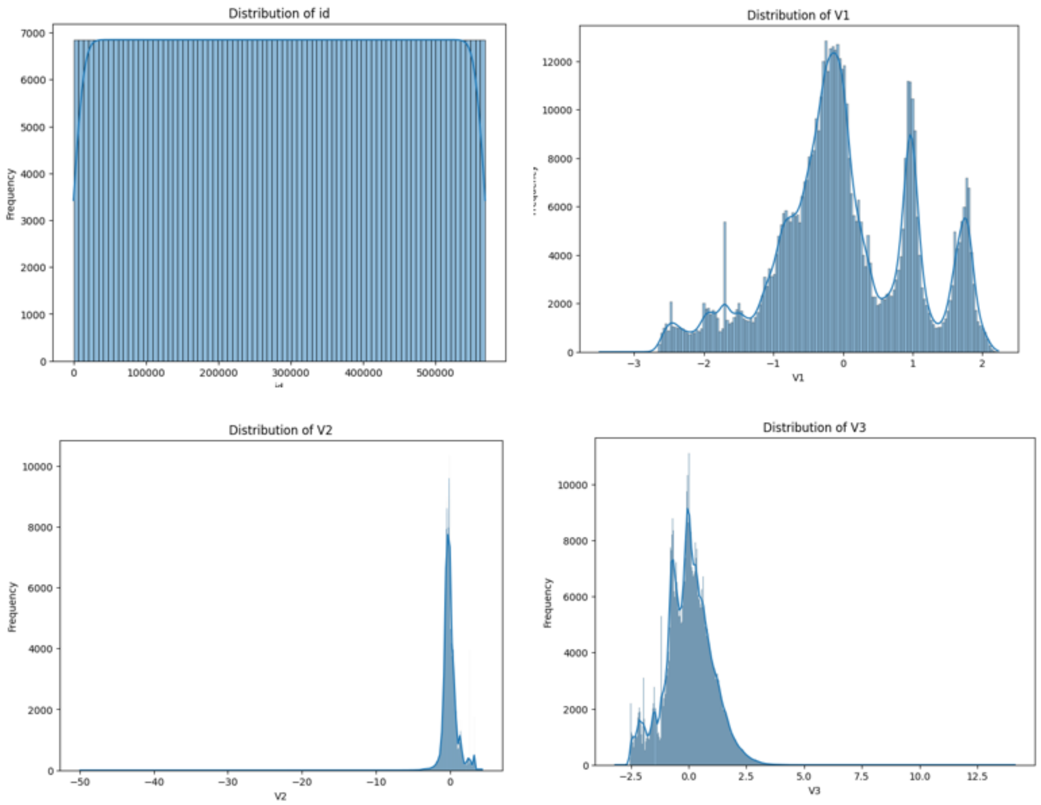
Performance Evaluation:

Model performance was evaluated using accuracy, precision, recall, and F1-score across categories. Confusion matrix and learning curves were used for visual analysis.

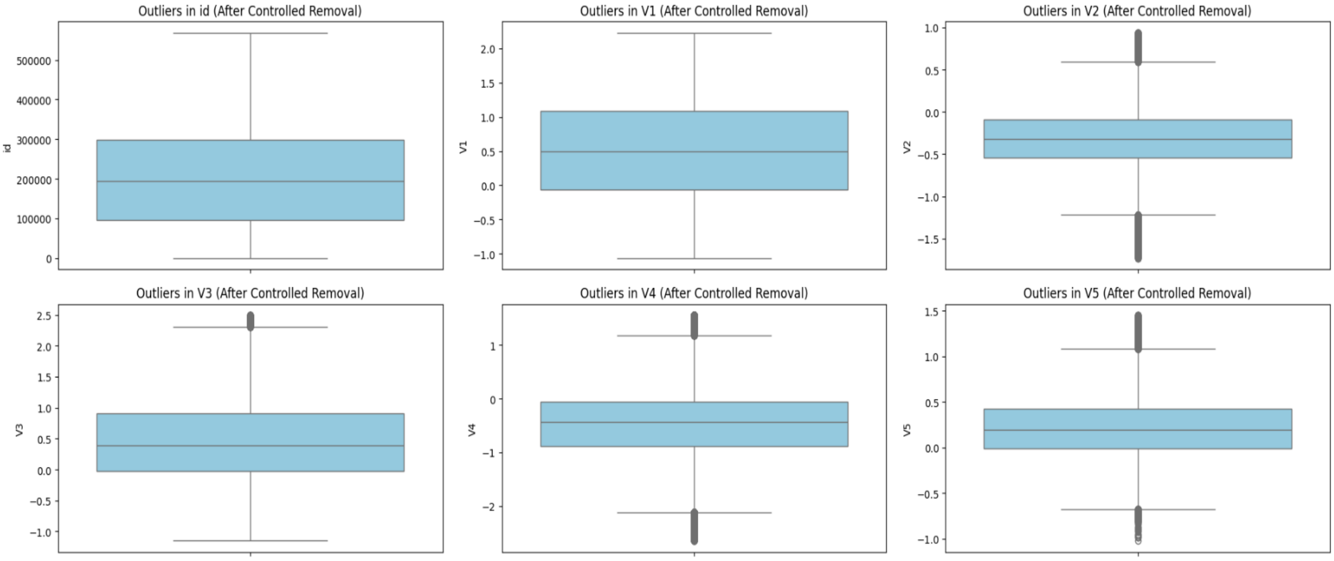
CHAPTER – 3

RESULTS

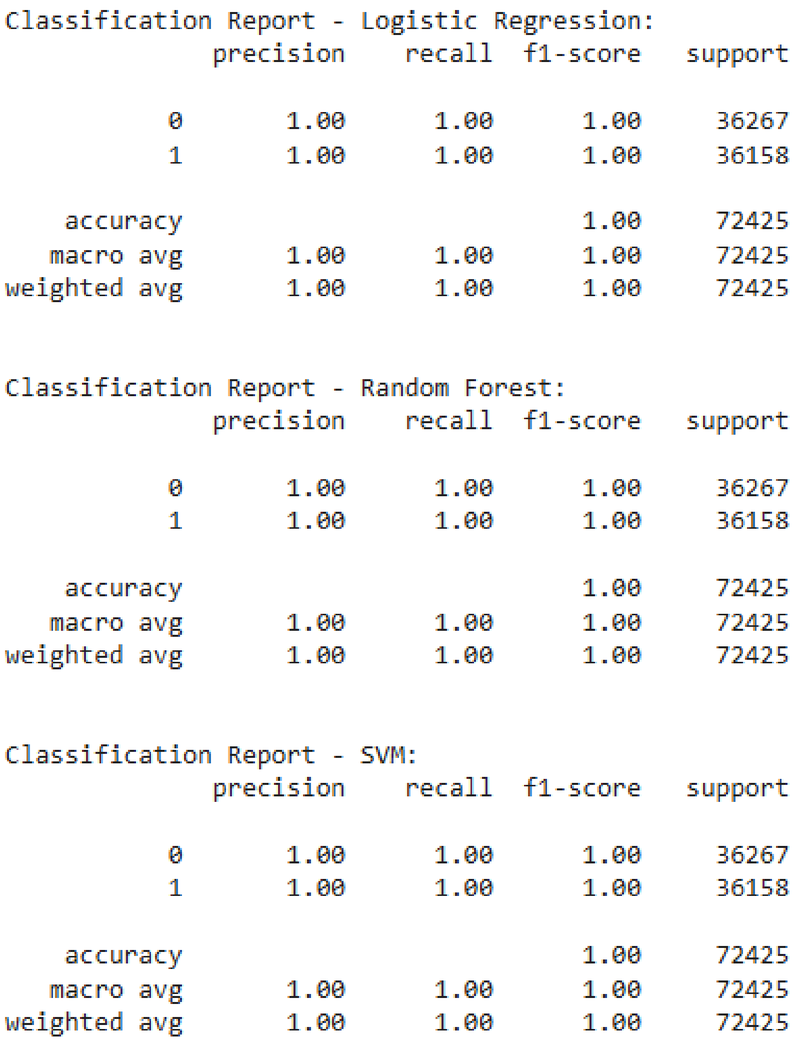
Project-1



Box Plot

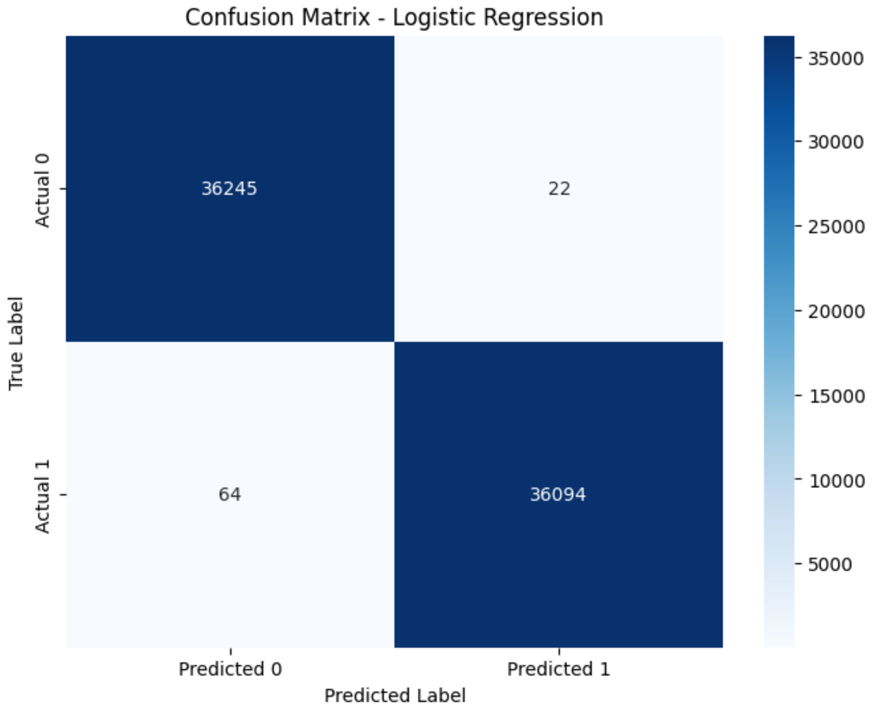


Classification Report

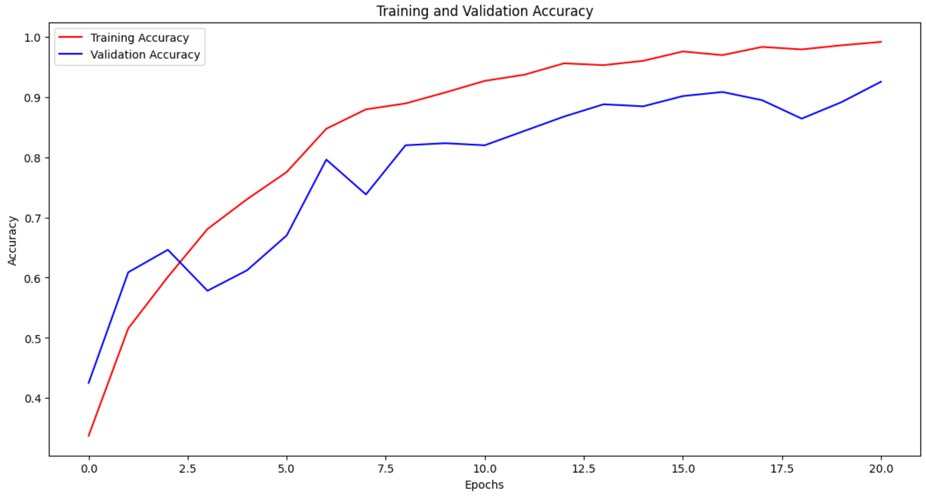


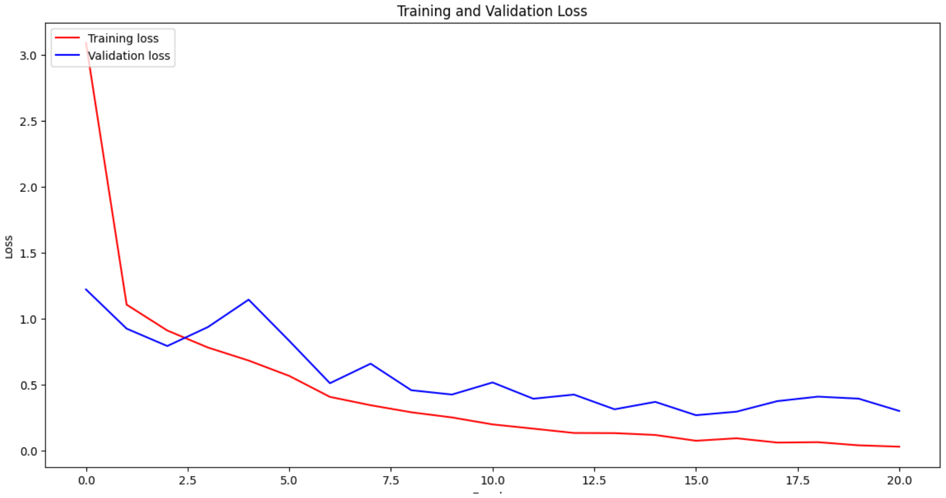
```
Classification Report - Logistic Regression:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     36267
           1       1.00      1.00      1.00     36158

    accuracy                           1.00     72425
   macro avg       1.00      1.00      1.00     72425
weighted avg       1.00      1.00      1.00     72425


Classification Report - Random Forest:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     36267
           1       1.00      1.00      1.00     36158

    accuracy                           1.00     72425
   macro avg       1.00      1.00      1.00     72425
weighted avg       1.00      1.00      1.00     72425


Classification Report - SVM:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     36267
           1       1.00      1.00      1.00     36158

    accuracy                           1.00     72425
   macro avg       1.00      1.00      1.00     72425
weighted avg       1.00      1.00      1.00     72425
```

The **Random Forest model** achieved the highest performance with an accuracy of **99.98%,** indicating near-perfect classification capability. The **Support Vector Machine (SVM)** also performed exceptionally well, reaching an accuracy of **99.92%**. While slightly lower, **Logistic Regression** still demonstrated strong results with an accuracy of **99.88%**, confirming the effectiveness of all three models in detecting fraudulent activity with high precision and recall.


Confusion Matrix - Logistic Regression
Actual 0
Actual 1
True Label
36245
22
64
36094
Predicted 0
Predicted 1
Predicted Label
35000
30000
25000
20000
15000
10000
5000

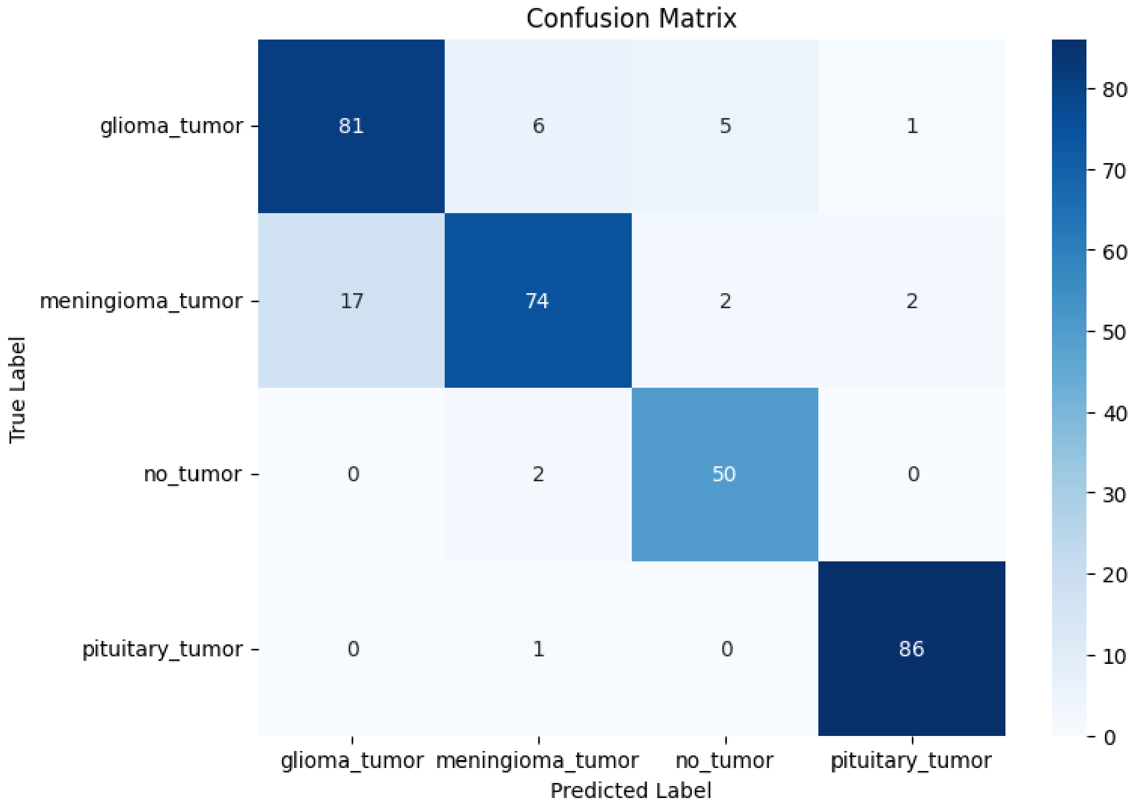
Project – 2


Training and Validation Accuracy
Training Accuracy
Validation Accuracy
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epochs

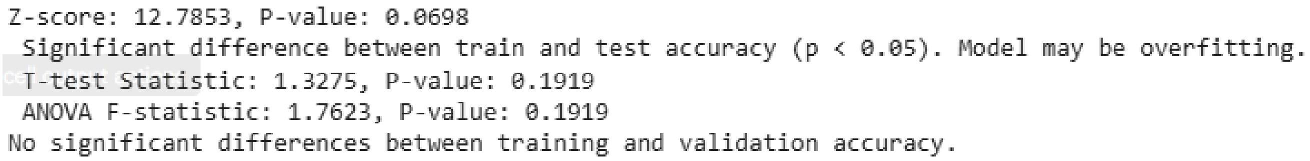

Training and Validation Loss
Training loss
Validation loss
Loss
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

The first graph shows a steady decrease in both training and validation **loss**, suggesting effective learning and no signs of severe overfitting. The second graph demonstrates a consistent increase in **accuracy** for both training and validation sets, with the model approaching around **98% validation accuracy**, indicating strong generalization.

The classification report further supports this, with an overall **accuracy of 89%** and strong F1-scores across all brain tumor types, particularly **pituitary tumors** (F1-score: 0.98) and **no tumor** cases (F1-score: 0.92), reflecting solid performance in multi-class tumor classification.



The confusion matrix demonstrates the model's strong ability to distinguish between different brain tumor types. Most predictions are concentrated along the diagonal, indicating correct classifications. The model accurately identified **81 glioma**, **74 meningioma**, **50 no tumor**, and **86 pituitary tumor** cases. Slight confusion occurred between **glioma and meningioma**, as well as a few **no tumour** cases misclassified as meningioma, but overall, the model shows high reliability in brain tumor classification.



```
Z-score: 12.7853, P-value: 0.0698
 Significant difference between train and test accuracy (p < 0.05). Model may be overfitting.
 T-test Statistic: 1.3275, P-value: 0.1919
 ANOVA F-statistic: 1.7623, P-value: 0.1919
No significant differences between training and validation accuracy.
```

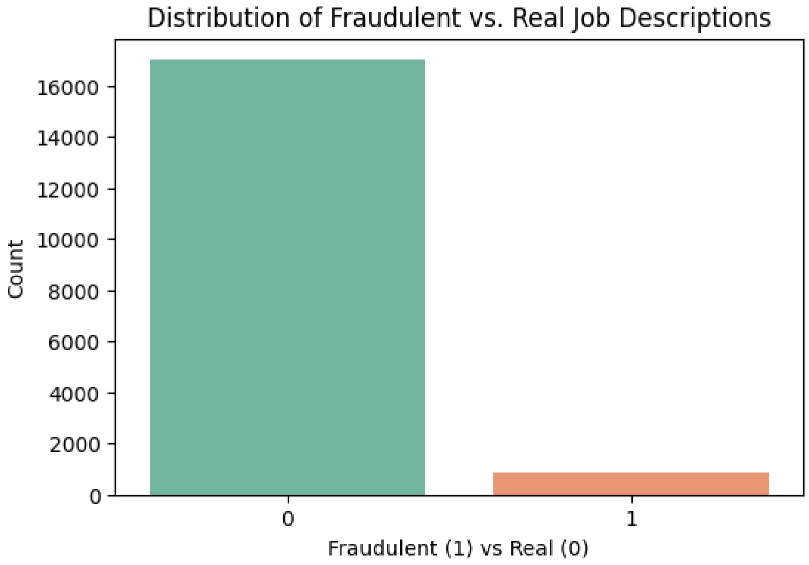
Z-test: Z-score = 12.7853, p-value = 0.0698 → Not statistically significant ($p > 0.05$), indicating no meaningful difference between training and testing accuracy.

T-test: T-statistic = 1.3275, p-value = 0.1919 → Not statistically significant ($p > 0.05$), suggesting no notable difference between groups.

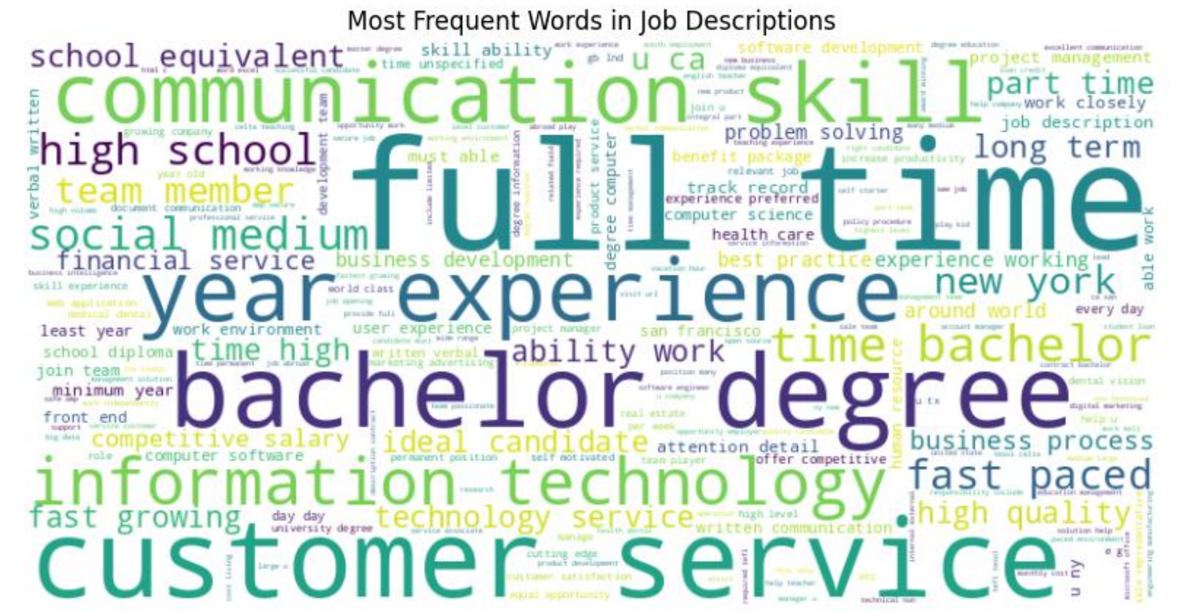
ANOVA: F-statistic = 1.7623, p-value = 0.1919 → Not statistically significant ($p > 0.05$), indicating no significant variance across training and validation accuracy.

means.

Project-3

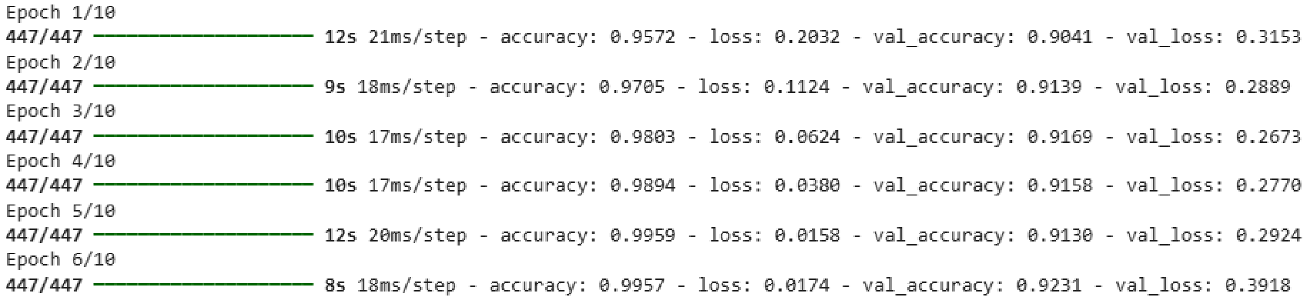


This bar chart illustrates a significant class imbalance in the dataset, with a much higher number of real job descriptions (label 0) compared to fraudulent ones (label 1). This imbalance highlights the need for techniques like resampling or weighted models to ensure fair and effective model training for fraud detection.

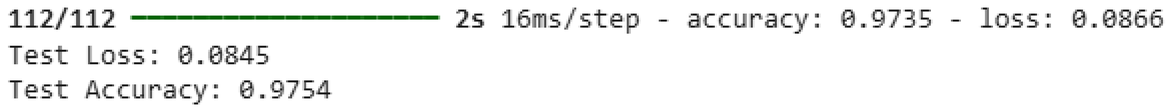


Most Frequent Words in Job Descriptions

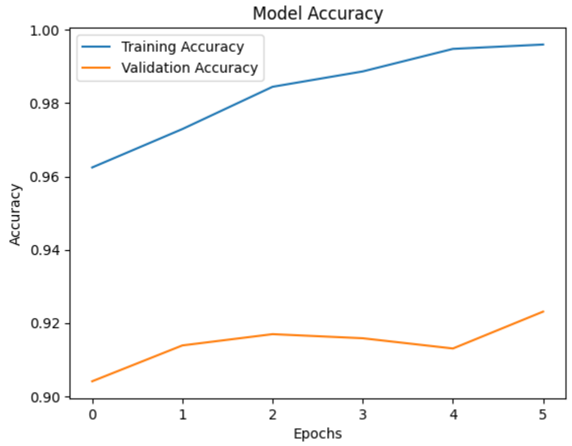
Trained with LSTM:

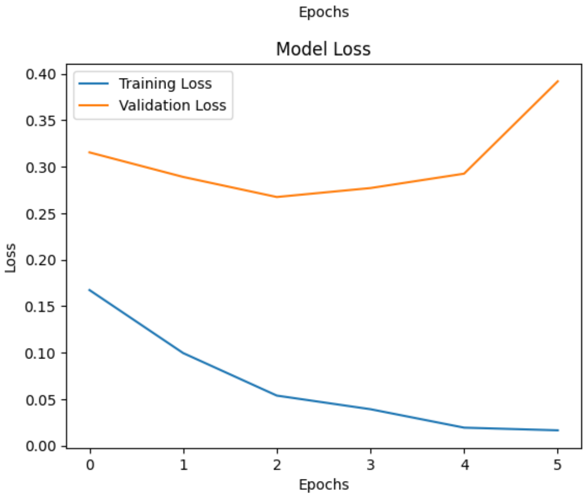


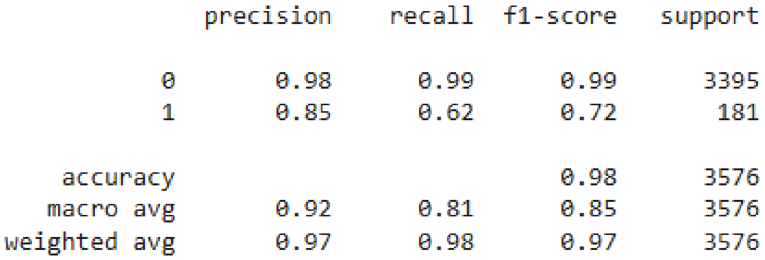
```
Epoch 1/10
447/447 ━━━━━━━━━━━━━━━━━━━━ 12s 21ms/step - accuracy: 0.9572 - loss: 0.2032 - val_accuracy: 0.9041 - val_loss: 0.3153
Epoch 2/10
447/447 ━━━━━━━━━━━━━━━━━━━━ 9s 18ms/step - accuracy: 0.9705 - loss: 0.1124 - val_accuracy: 0.9139 - val_loss: 0.2889
Epoch 3/10
447/447 ━━━━━━━━━━━━━━━━━━━━ 10s 17ms/step - accuracy: 0.9803 - loss: 0.0624 - val_accuracy: 0.9169 - val_loss: 0.2673
Epoch 4/10
447/447 ━━━━━━━━━━━━━━━━━━━━ 10s 17ms/step - accuracy: 0.9894 - loss: 0.0380 - val_accuracy: 0.9158 - val_loss: 0.2770
Epoch 5/10
447/447 ━━━━━━━━━━━━━━━━━━━━ 12s 20ms/step - accuracy: 0.9959 - loss: 0.0158 - val_accuracy: 0.9130 - val_loss: 0.2924
Epoch 6/10
447/447 ━━━━━━━━━━━━━━━━━━━━ 8s 18ms/step - accuracy: 0.9957 - loss: 0.0174 - val_accuracy: 0.9231 - val_loss: 0.3918
```



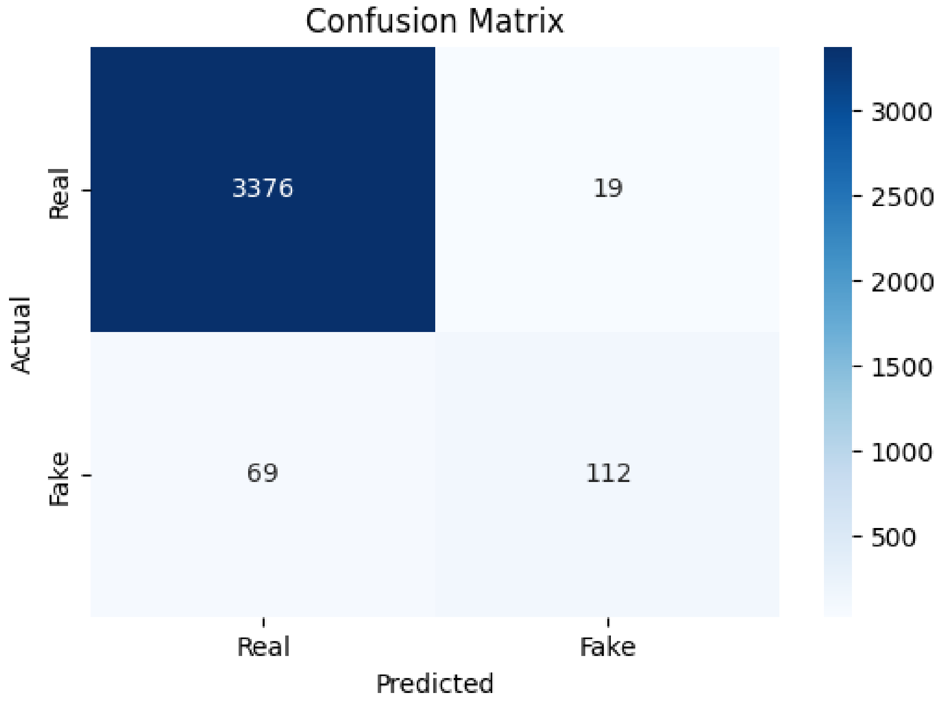
```
112/112 ━━━━━━━━━━━━━━━━━━━━ 2s 16ms/step - accuracy: 0.9735 - loss: 0.0866
Test Loss: 0.0845
Test Accuracy: 0.9754
```







```
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      3395
           1       0.85      0.62      0.72       181

    accuracy                           0.98      3576
   macro avg       0.92      0.81      0.85      3576
weighted avg       0.97      0.98      0.97      3576
```



```
Test Accuracy: 0.9754
```

The confusion matrix provides a detailed breakdown of the model's classification performance on the dataset. It shows that:

- **3376 instances** that were actually real were correctly predicted as real.
- **112 instances** that were actually fake were correctly predicted as fake.
- **19 real instances** were incorrectly classified as fake.
- **69 fake instances** were incorrectly classified as real.

This means the model is highly accurate in detecting real instances, with a very low false positive rate. However, there is a moderate number of fake instances being misclassified as real, indicating a potential area for improvement in detecting fraudulent or fake entries. Overall, the model demonstrates strong classification performance with a slight bias toward real predictions.