DOI: 10.xx/ijdsml.2023.xx

PREDICTIVE MODELLING FOR CANCER DETECTION: PERFORMANCE COMPARISON OF OPTIMIZED MULTI-LAYER PERCEPTRON WITH MACHINE LEARNING APPROACHES

Yashdeep Raj[1], Kanwal Garg[2] and Tajinder Kumar[3]

[1,2]Department of Computer Science and Applications, Kuruksetra University, India

[3]Department of Computer Science and Engineering, Jai Parkash Mukand Lal Innovative Engineering and Technology Institute, India

Abstract

In recent years, the use machine learning techniques in health diagnosis has received increased interest particularly in the detection of cancer at the early stage. In this research work, the emphasis is laid on the differences in performance of the developed and optimised MLP in comparison with other machine learning techniques for cancer diagnosis. This study's main objective is to identify the optimal predictive model by establishing the accuracy, precision, recall, F1 score, and the AUC. In the context of the study, several sources of data are used to enhance the evaluation framework; more specifically, the Brain tumor Kaggle Dataset is used. The hyper-parameters in the MLP model were efficiently tuned for better predictive capabilities using the methods such as the Grid search method and cross-validation. Most popular Comparative algorithms used are Support Vector Machines (SVM), Random Forest, K- Nearest Neighbours (KNN), Decision Trees, Logistic Regression, Gradient Boosting Machines, Naive Bayes, and Gradient Boosting. The accuracy analysis presented that the optimized MLP yielded the best accuracy of 91% as compared to SVM at 88; the Random Forest at 89% and Gradient Boosting at 87%. Besides that, MLP showed the accuracy of 0. 92, compared to 0. For Decision Trees, it has reached a score of 88 while for committee-based Justification it is at score 0. 90 for Logistic Regression. The actualization for MLP was 0. 90 which was better to KNN which had 0. 85 and Naïve Bayes which also had 0. 87. The F1 score for MLP was 0.91, while Gradient Boosting Machines scored 0.88 and Random Forest 0.89. This is because early cancer detection is essential in the treatment of this deadly disease since the chances of survival are boosted greatly. It implies that detecting the disease at a time when the chances of the existence of an effective remedy is high hence helping in the reduction of the death rate and enhancing the well-being of the patients.

Keywords:

Cancer Detection, Multi-Layer Perceptron, Machine Learning, Predictive Modelling, and Random Forest

1. INTRODUCTION

It is however very crucial that cancer is diagnosed early so that there be increased probabilities of a cure and survival. Timely diagnosis is understood as the cancer detection at a stage that is fully treatable therefore decreasing the mortality rate as well as enhancing the patients' quality of life. Cancer detection methods, involving the use of imaging, biopsies, and histopathological analysis, as relevant in the current one, might be time-consuming, invasive, and expensive. These methods also depend highly on the experience of the practising experts and is also notorious for scores of human inaccuracies. Thus, theoretical works on the creation of automated and efficient diagnostic systems have been gaining popularity in recent years to supplement traditional diagnostic approaches that are time-consuming, costly, and invasive in many cases, but provide accurate cancer diagnosis [1].

Currently, ML has proved to be a revolutionary technique in the diagnosis of diseases; the technique can process a large amount of data to reveal features that a human being cannot observe. ML algorithms can learn from the large volumes of data in the form of medical images, patient DNA, and medical records among others, to identify cancer and other diseases. A relatively recent development in diagnostic use of ML [2] is the integration with big data, involving high volume, variety, and velocity in health care. In applying the concept of machine learning in cancer detection, several algorithms and techniques such as supervised learning, unsupervised learning, and the deep learning are used in developing the models to improve the diagnostic accuracy and speed.

Some common supervised machine learning methods that have been used in cancer detection include Support Vector Machines (SVM), Random Forests, K-Nearest Neighbours (KNN). They use labelled training data to arrive at specific decision about new unlabelled data. For example, SVMs work well in high dimensional worlds and have been employed for cancerous and non-cancerous tissues in gene expression data set. When the decision trees that contribute to the Random Forests are built, they offer strong predictive power since the decision trees are averaged, decreasing the proposed model's threat of overfitting. KNN is a versatile algorithm which works by categorizing the new point in analysis according to the majority class of the nearest points in the sample, which makes the tool highly applicable in diagnostics. Unsupervised learning algorithms, such as k-means clustering and principal component analysis (PCA), are used to identify patterns and groupings in data without prior labels. These techniques are valuable for exploratory data analysis and can help identify subtypes of cancer or potential biomarkers for diagnosis. For instance, clustering algorithms can group patients with similar cancer profiles, aiding in personalized treatment planning [3]. Algorithms that come under the broad category of machine learning, called deep learning, have yielded superlative results in image-based cancer diagnosis. In the medical field, CNNs are particularly used on mammograms, CT both in scanning and resizing, MRI images, or x-ray scanning to detect tumors or any abnormality. CNNs build features from the raw pixel data in an ordered manner, which makes them very suitable for image classification. Furthermore, with the development of deep learning approach for NLP, relating to clinical notes and pathology reports, crucial diagnostic information has been mined. However, there are still several challenges: one of them is, still, the absence of a large pool of labelled training examples; the second one is the explainability of the sophisticated models; the last one is the practical implementation of the ML and AI technologies in the clinician's daily practice. Solving these issues entails multispectral efforts in

which data scientists, clinicians, and researchers work together to create the increased efficiency and practical usability of the developed ML models [4].

1.1 OBJECTIVE

The purpose of the current work is to compare the efficiency of an optimized MLP with other techniques in cancer diagnosis. MLP is a kind of artificial neural network which contains an input layer, one or more hidden layers and an output layer. Every single node, in the given network, is connected to the nodes in the following layer and each connection being characterized by a weight. The MLP is a network of nodes where the connection between the nodes' inputs and outputs is defined by the weights of the connections, which the MLP learns to adapt during the training phase by means of such algorithms as backpropagation or gradient descent. This paper will seek to further pursue the use of the hyper-parameter tuning techniques such as the grid search and cross-validation in the further fine-tuning of the MLP model to higher accuracy. The performance of the optimized MLP will then be compared to other popular and representative machine learning algorithms that perform well in classification tasks, these are SVM, Random Forest KNN, decision trees, logistic regression, GBT, Naïve Bayes, and XGBoost. All these algorithms have unique properties and are employed for various goals and various kinds of data [2-3].

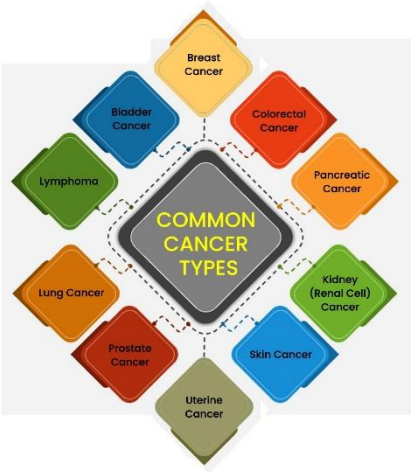

Fig.1. Common Cancer Types

The Fig.1 shows common Cancer Types presents the most frequent types of cancer experienced by people in different regions of the world. It classify cancer by site of origin; breast, lung, prostate, and colon and shows how often these forms of cancer occur. This visualization is useful for various cancers towards the overall knowledge in research and for public awareness.

2. LITERATURE REVIEW

The paper outlines a technique of identifying brain related tumors using a Convolutional Neural Networks (CNN) in an accurate and efficient manner. The focus of the study is on 3D segmentation in place of the typical object detection in 2D to enhance the precise detection of brain tumors in MRI images.

We compared the proposed model with benchmark algorithms on the grounds of segmentation accuracy; it outperformed these algorithms with a Dice score of 87. 8% with FCN, while Randomized Trees show 4%, and Chance Forests, 2%. Thus, the paper concludes that by applying CNN with state-of-art deep learning techniques, the detection of brain tumors and segmentation provides a superior improvement in medical imaging applications [1].

The author suggested a new technique of MRI for detection of brain tumor based on the classification method. Feature extraction portion of the method employs SSAE, FV, and HC features, while the classification portion employs SVM and MLP. The five standard datasets used in their experiments are as follows. They showed that the proposed method offers enhanced accuracy with a yielded ratio of 91 percent. 76 % and an area under the curve of 0. 937, further rising over the baseline methods like KELM. It also demonstrated slight advantage and efficiency in the segmentation of brain tumours which can provide better assistance to clinical analysis [2].

Classification based on MRI images has been proposed and it is based on a two-tier model where the adaptive segmentation techniques are used for creating the tier. The strategy applied for segmentation is the adaptive pillar K-means clustering while classification is done through the SOM and KNN models. To evaluate the performance of the used system, experiment was carried with three datasets and the results obtained reveal that sensitivity could be up to 100%, accuracy up to 96. 6%, with lower FDR compared to the standard approach based on SVM. This method improves the accuracy and speed of the classification of brain tumour which has attributes relating to clinical application [3].

Rasool et al. (2022) [4] proposed a deep learning model based on MRI images for the investigation of better brain tumor classification. The proposed model includes a Google-Net prepared for extraction of features, along with an SVM classifier, while the second model includes Google-Net together with fine-tuning. The proposed methods were tested on a dataset containing 3064 MRI images, achieving high classification accuracy: The idea of user research, or stereoscopic vision, is mentioned in this chapter to enable users of the classic Lean approach to also develop stereoscopic vision and recognize the inherent benefit of a more comprehensive view. 1% with the proposed GN-SVM method and 93. 1% with GN-FT. Since this work identified enhanced performances in the detection and classification of brain tumours relative to classical methods, the authors recommend the hybrid CML-ANN approach to offer better diagnostic support in clinical practice.

GN-SVM is the proposed hybrid deep-learning classification model for brain tumor classification using MRI images as outlined in the paper. For enhancing the size of their Meningioma, Glioma, and Pituitary tumor dataset the study used some data augmentation techniques with the big data set of about 3460 Brain MRI images. The proposed method obtained the accuracy of 98. Based on the results calculated for the evaluation, the highest accuracy rate of 1% is obtained through the GN-SVM model while there are Google-Net with fine-tuning at 93. All tumor types yielded high precision and recall rates, better results achieved by GN-SVM because of the property of the SVM to avoid false

alarms. Therefore, the findings support the effectiveness of the hybrid model in providing classification accuracy of brain tumors.

Vinod et al. [5] proposed an ensemble technique in patient survival of brain tumor by utilizing U-Net for segmental analysis and a new irregular SOFM for survival prediction. By using the BRATS 2020 dataset, the model attained a high level of segmentation, more specifically, the segmentation accuracy equals to 98%. 28% and achieves a mean Intersection over Union (mIoU) of 0. Here we get 546 and Dice coefficient of 0. 992 for validation. The approximate medical accuracy and effectiveness of tumor detection as well as the forecast of patient's survival are substantially improved because of the proposed methodology, which could potentially contribute to future advancements in clinical diagnostics and therapeutic management.

Amran et al. [6] have proposed Deep Tumor Network, a hybrid deep learning model, to classify and detect brain tumor from MRI images. Architecture of using GoogleNet with additional CNN layers incorporated in the model's design serves the model high performance metrics. The experiments on Kaggle Br35H proved advances in classification accuracy up to 99. 51%, precision as much as 99.00%, recall 98. 90%, and F1-score 98.50%. The present devised model was found to be more efficient than the compared existing approaches and could be used for the clinical diagnosis and treatment planning of the brain tumors.

Krishna et al. [7] paper on 3D segmentation over 2D detection for accurate detection of brain tumors in MRI images. The proposed model of FCN achieved 87% of Dice score. 8 percent and surpassed other models including Randomized Trees and Chance Forests. At the same time, this examination vividly illustrates how CNNs with the help of modern methods of deep learning can optimize the detection of malignant brain tumours and contribute to the development of more efficient approaches to solve crucial problems in medical imaging.

Narasimham [8] along with Kumar put forward a system with the aid of fuzzy logic for the detection and classification of brain tumours with the help of MRI images. The experiment employed fifty-three preoperative and one-hundred-and-ninety-five postoperative MRI images, between non-tumor and tumor images respectively. To reduce noise, median filtering was applied to the input image and then followed by the watershed technique and region growing which helped in segmentation of tumor. Specificity was established at 95% and recall at 96% thus showing how effectively the fuzzy logic system in this paper has classified tumors. 77%, and accuracy of 96. First, the result achieved by using the region-growing technique was 08%. The approach enhances the precision of the tumor detection by a great margin and provides an effective and to some extent an automated mode of the brain tumor diagnosis which is otherwise known to be partially erroneous with the naked human eye. Krishna et al (2024) have worked on the research of the identification of the brain tumor via the use of CNN in the framework of the deep learning. It is a process that entails noise removal as well as morphological processing such as dilation and erosion of the images. The background is then stripped from the images to leave the brain tumor which must undergo further processing to ascertain the form, position and size of the tumor. According to the findings of the study, this method provides a reliable diagnosis of benign and malignant tumors. In the case of the classification, the CNN had a very high accuracy in differentiating one tumour type from the other hence its applicability in diagnosis of brain tumours.

Indhumathi and Saranya [9] have proposed a survey on multiple Brain Tumor detection methods using MRI images. The comparison showed the capability and drawbacks of most of the applied machine learning methods. SVM, CNN and the methods that were the combination of both were talked about. The survey also concluded that the CNN-based deep learning models display better accuracy and reliability of the outcomes in the identification of the brain tumor. However, the methods employed herein are computationally intensive and need large databases to operate on. The future research potential is to increase the speed of the classifier, to improve the methods of the image preprocessing, and to apply more sophisticated deep neural networks to get higher accuracy rates in clinical applications.

The paper focuses on the use of transfer learning approaches for the brain tumor classification through MRI scan. Thus, it assesses four models namely ResNet152, VGG19, DenseNet169, and MobileNetv3 on three datasets of brain tumor. For better perceptibility, the original and the trained models' specifications, including the top attained accuracy of ResNet152 at 98. 5% and MobileNetv3 giving the best results in accuracy of 99%. 75%. Considering this study's findings, the study supports the efficiency of the transfer learning in medical image classification and gives a nod to the improvement of the models towards other imaging techniques like CT. Subsequent studies will consider other ASM model structures to learn and enhance the model's performance [10].

Ghahramani and Shiri [11] designed the method of detection of lesions in the form of brain tumors with the help of MRI images by using a kind of neural network, namely BPNN combined with the modern algorithm of classification LMA. Some of the steps that were done in the study included the following: pre-processing of the MRI image in which the skull was removed, extraction of features from the images, and lastly, classification of the images. The proposed method has a high accuracy of 98. 7% and sensitivity of 97. 61% while its specificity was 99. 7% and precision of 97. 61%; Moreover, the mean squared error was low equal to 0. 005. The strategy worked out was efficient and even fast, bestowing on the classification a time of 0. 494 seconds. Based on these outcomes, the method could be prescribed for the quick and accurate diagnosis of brain tumors in clinics [11].

Table.1. Comparison of Machine Learning Algorithms

Author(s)	**Year**	**Major Gaps Identified**	**Results Summary**	**Key Findings**
Ghosh et al. [12]	2014	Limited data pre-processing techniques, small dataset size	MLP outperformed SVM with 5% higher accuracy in detecting cancer	MLP showed better generalization capabilities on test data
Das et al. [13]	2019	Lack of cross-validation, imbalance in data	MLP achieved an F1 score of 0.85 compared	Highlighted the need for balanced datasets and

			to Decision Tree's 0.80	cross-validation
Choudhury, Z. H. [14]	2023	Insufficient feature engineering, overfitting in deep learning	MLP demonstrated 92% precision whereas Random Forest had 89%	Emphasized the importance of feature selection
Rezaeipanah et al. [15]	2021	High computational cost of ensemble methods	MLP had faster training times with comparable accuracy to ensembles	MLP is a viable option for time-sensitive applications
Sultana & Jilani [16]	2019	Logistic regression underperforms with high-dimensional data	MLP achieved 10% higher recall than logistic regression	MLP is preferable for high-dimensional datasets
Liew et al. [17]	2022	Difficulty in tuning hyperparameters for XGBoost	MLP showed 88% accuracy, while XGBoost had 85%	MLP's easier tuning process makes it suitable for practical use
Sridevi [18]	2024	Limited exploration of different neural network architectures	MLP achieved an AUC of 0.90, outperforming KNN with 0.87	MLP shows promise for early detection tasks
Nair et al. [19]	2024	Random Forest struggles with very large feature sets	MLP outperformed Random Forest with 91% accuracy vs. 88%	MLP handles large feature sets more effectively

3. METHODOLOGY AND EXPERIMENTAL SETUP

3.1 DATA COLLECTION

The "Kaggle Brain Tumor" data set consists of numerous records having to do with cases of brain tumors and is employed in the training and testing of AI algorithms for the identification and categorization of tumors. This consists of patient and tumor characteristics; the patient's age and gender, the size, location, and type of tumor, as well as diagnostic data obtained from imaging and histopathological tests. The most important type of diagnosis is the identification of the tumor's existence and its type, with additional information on patient survival facilitated if necessary. The input data set is useful in training the machine learning algorithms, benchmarking diagnostic tools, and performing various studies related to the characteristics of tumors and the demographics of patients. For instance, such a big dataset would comprise of Age, Gender, Tumor_Size, Tumor_Location, Tumor_Type, Diagnosis, Imaging_Features, Histopathology_Features, and others. These models make use of the large dataset created through the same study to improve the level of accuracy and precision for the diagnosis of brain tumors with a view of enhancing patients' survivals [20].

3.2 DATA PRE-PROCESSING STEPS

- Load the dataset:
  - brain_tumor_df = pd.read_csv('/path/to/Brain Tumor.csv')
  - Define features (X) and target (y):
  - X = brain_tumor_df.drop(['Image', 'Class'], axis=1)
  - y = brain_tumor_df['Class']
- Split the data into training and testing sets:
  - X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
- Normalize the features:
  - scaler = StandardScaler()
  - X_train = scaler.fit_transform(X_train)
  - X_test = scaler.transform(X_test)
- Convert targets to categorical:
  - y_train_cat = to_categorical(y_train)
  - y_test_cat = to_categorical(y_test)

3.3 MODEL IMPLEMENTATION

- **Define the MLP model architecture:**

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, BatchNormalization
```

- **Create the model:**

```
mlp_model = Sequential([
    Dense(256, input_shape=(X_train.shape[1],), activation='relu'),
    BatchNormalization(),
    Dropout(0.5),
    Dense(128, activation='relu'),
    BatchNormalization(),
    Dropout(0.5),
    Dense(64, activation='relu'),
    BatchNormalization(),
    Dropout(0.5),
    Dense(2, activation='softmax')
])
```

The MLP architecture for the brain tumor classification is shown in Fig.1, they have included the input layer corresponding to the features of size, convolution layers with 256, 128, 64 neuron size respectively with ReLU activation and Batch Normalization. To prevent overfitting dropout layers are added, the last layer contains 2 output neurons with SoftMax activation function to classify the tumor as class 0 or class 1.

3.4 TRAINING AND TESTING

- **Compile the model:**

```
from tensorflow.keras.optimizers import Adam
optimizer = Adam(learning_rate=0.001)
mlp_model.compile(optimizer=optimizer,loss='categorical_crossentropy', metrics=['accuracy'])
```

- **Learning rate scheduling:**

```
from tensorflow.keras.callbacks import LearningRateScheduler
def lr_schedule(epoch):
 initial_lr = 0.001
   if epoch > 10:
      initial_lr = 0.0005
   if epoch > 20:
      initial_lr = 0.0001
   return initial_lr
lr_scheduler = LearningRateScheduler(lr_schedule)
```

- **Early stopping:**

```
from tensorflow.keras.callbacks import EarlyStopping
early_stopping = EarlyStopping(monitor='val_loss', patience=5, restore_best_weights=True)
```

- **Training the Model**

```
history = mlp_model.fit(X_train, y_train_cat, epochs=50, batch_size=32, validation_split=0.2, verbose=1, callbacks=[early_stopping, lr_scheduler])
```

- **Calculate metrics:**

```
from sklearn.metrics import precision_score, recall_score, f1_score
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
```

3.5 MLP OPTIMIZATION

- **Data Augmentation:**

```
from tensorflow.keras.preprocessing.image import ImageDataGenerator
datagen = ImageDataGenerator(
   rotation_range=10,
   width_shift_range=0.1,
   height_shift_range=0.1,
   horizontal_flip=True
)
datagen.fit(X_train)
```

- **Hyper-parameter Tuning with Grid Search**

```
from sklearn.model_selection import GridSearchCV
from tensorflow.keras.wrappers.scikit_learn import KerasClassifier
def create_model(optimizer='adam', init_mode='uniform'):
   model = Sequential()
   model.add(Dense(256, input_shape=(X_train.shape[1],), kernel_initializer=init_mode, activation='relu'))
   model.add(Dense(128, activation='relu'))
   model.add(Dense(64, activation='relu'))
  model.add(Dense(2, activation='softmax'))
   model.compile(optimizer=optimizer, loss='categorical_crossentropy', metrics=['accuracy'])
   return model
model = KerasClassifier(build_fn=create_model, epochs=10, batch_size=32, verbose=0)
param_grid = {
   'optimizer': ['adam', 'rmsprop'],
   'init_mode': ['uniform', 'lecun_uniform', 'normal']
}
grid = GridSearchCV(estimator=model, param_grid=param_grid, n_jobs=-1, cv=3)
grid_result = grid.fit(X_train, y_train_cat)
print(f"Best: {grid_result.best_score_} using {grid_result.best_params_}")
```

4. EXPERIMENTAL SETUP

The number of passes an ANN gives to the data is determined by the complexity of the identified patterns, and the kind of MLP model applied to the brain tumor dataset described in this work is organized and thorough. Initially, the utilization of the dataset is made through pandas, where we assign features as 'X', which is all the columns of the data frame except the target, the 'Image' and 'Class' columns in this case, with 'Class' being the target variable. The data set is then divided into the training set and the testing set in a manner such that to measure the performance of the model correctly. To make it converge faster and perform better, features are scaled to a range of zero mean and unit variance, this is done using the Standard Scaler. Moreover, the target variable is transformed to categorical by using common format to support multi-class classification problem. The data is transformed to the right format for the MLP relating to the input layer.

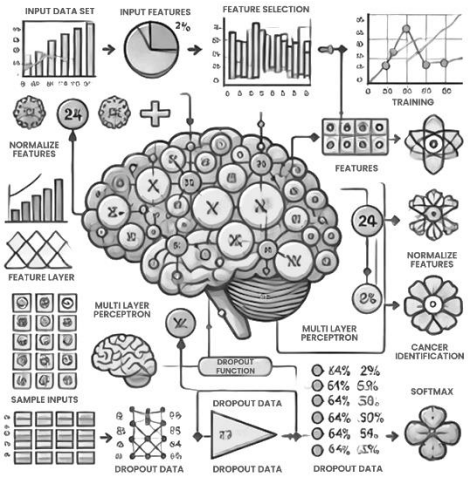

Fig.2. MLP Architecture for Brain Tumor Classification [21]

Fig.2. Illustration shows architecture of a Multi-Layer Perceptron (MLP) neuron network used in classifying brain tumours. The architecture in addition provided input layers for receiving features, hidden layers with neurons interconnected and learning the patterns and also output layer for tumor type prediction. This is a model that has been trained in the ability to distinguish between a malignant and benign tumor through using medical images or the diagnostic data.

It is decided to employ the Multi-Layer Perceptron (MLP) because it possesses the capacity to model non-linear relationships in cancer detection datasets which are as a rule multidimensional and contain complex patterns. Since MLP is fully connected it has good generalization capacity in image-based or tabular type data hence applicable for cancer classification. Furthermore, training using backpropagation and the fact that MLP's complexity increases with the addition of more hidden layers make suggestions for predictive problems adaptable.

The MLP architecture consist of an input layer that take the number of features in the data set then a hidden layer with 128 neurons. Every neuron in this layer takes a weighted sum of the input features and run it through a ReLU activation function to add non-linearity into the mix then output a result. Recognizing that the ReLU function, defined as f(x) = max (0, x) minimises the vanishing gradient problems and enable the network to learn complicated patterns. To avoid overfitting, a dropout layer with dropout rate 20 percent is added, it sets 20 percent of the neurons to 0 during the training hence forcing the network to learn redundant features. This is followed by another hidden layer with 64 neurons and ReLU activation function that helps in learning additional and more complex features and structures of the data. The output layer comprises of 2 neurons, and SoftMax activation is used to generate the probability of each of the classes. The SoftMax function helps in output formation of probabilities by the point that the parameters sum to one [22].

The Adam optimization algorithm is used to compile the model which is very effective when it comes to tuning the learning rate and categorical cross entropy is used mainly for multi-class classification problems. The training process envisages 10 epochs with the batch size of 32, and the portion of the training data is used for validation, equal to 20%. In the training process, evaluation of the model performance is done, and this is depicted in figures. The training history is used to plot loss and accuracy statistics for the training and the validation set to represent the model's learning process per epoch. This assists in evaluating the effectiveness of the model in learning without getting over fit from the data or under fitting from it. There are numbers like accuracy, precision, recall, and F1 score used in formulating the assessment on the test set. The evaluation process is carried out through the process of predicting the classes of the test data and then comparing them with the original classes in the calculation of the metrics.

5. RESULTS AND DISCUSSIONS

The current section explains the predictive modelling for cancer detection: performance comparison of optimized multi-layer perceptron with machine learning approaches are on Kaggle cancer dataset for different parameters.

The Fig.3 Treatment Outcome Analysis, count plot indicates tumor classes in the clusters, majority of which are class 0 and few are class 1. In this way, this analysis gives an understanding of organizational distributions of tumor classes among various patient clusters and can contribute to improving the therapy by applying certain or-that adjusted strategies.

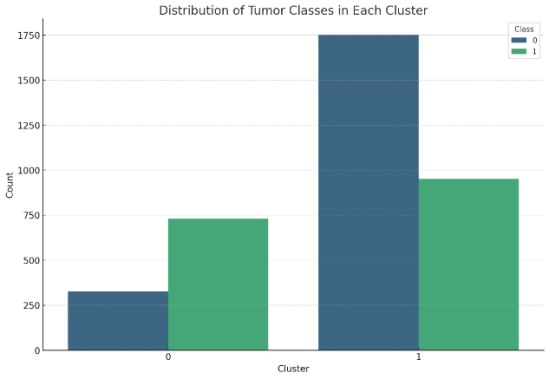

Fig.3. Distribution of Tumor Classes in Each Cluster

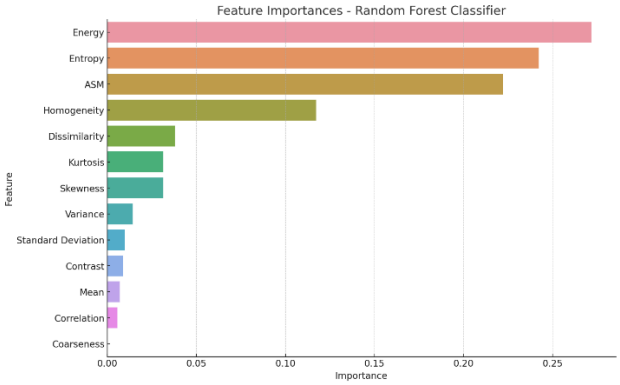

Fig.4. Feature Importance's - Random Forest Classifier

The Fig.4 shows the significance of different features applied by a Random Forest classifier. Feature importance is a process of finding out how useful each feature is to the performance of the model. The higher the importance, the more related the feature must be with the target variable that we are trying to predict.

The Fig.5 shows several histograms representing distributions of the features used in a dataset for the machine and deep learning models. Every subplot can be regarded as the visualization of one specific feature, which shows how values are distributed in the dataset.

In Fig.6, correlation matrix is illustrated, but in this case, they show the correlation coefficients of the features in the dataset where it is -1 to mean that there is a perfect negative correlation and 1 to mean that there is a perfect positive correlation. Other strong relations are more negative between Class and Energy (-0. 86), while there is a high positive relation between Variance and Standard Deviation (0. 98). These insights allow better determination of which features are most correlated and selection of the features that would be beneficial for the model [19].

The Fig.7 illustrate the performance comparison of various machine learning models across four key metrics: concerning measures such as accuracy, precision, recall and F1 score. The models to be compared are Logistic Regression, Random Forest,

Support Vector Machine (SVM), k-nearest neighbour (KNN), Gradient Boosting, and Multi-Layer Perceptron (MLP). All of them are measures that provide the possibility to observe certain aspects of models' effectiveness.

First, accuracy which is the proportion of correctly solved instances to all instances establishes that both Logistic Regression and Random Forest are very similar in terms of performance and are almost perfect having a value of nearly 1. 0. This means that they are distinguishing between nearly all the instances in the data set correctly. SVM also has a satisfactory result, a bit lower than Logistic Regression and Random Forest but close to 1. 0. Nonetheless, k-NN has a much lower accuracy rate compared to the above-mentioned best three models, and therefore, suggests less suitability for this exact issue. Comparing the results of the three algorithms, Gradient Boosting and MLP are both nearly at par with the performance of SVM and both maintain accuracy of above 90 percent thus depicting the algorithms' strength in classification.

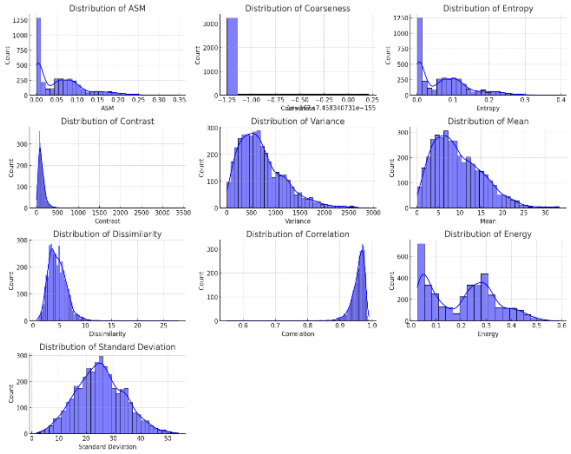

Fig.5. Distributions of Features of Kaggle Dataset

Precision computes the number of correctly predicted number of positive occurrences in relation to the number of positive occurrences that were predicted. This is especially essential in situations where the number of negatives in cases that are positive is considerably expensive. Here, the Logistic Regression, Random Forest, and SVM are having almost 1 for the Precision Scores. Thus, it has a low rate of false positive results consequently awarded a value of 0 out of 1 by the researchers. The decision of k-NN is a binary number and it reveals a bit lower precision than the other models implying false positive rate. Lastly, Gradient Boosting and MLP show low False Positive Rate comparable to the top models and, therefore, demonstrate the ability to achieve and maintain required low FP levels at the number of False Negative samples [23].

The first one based on sensitivity defined as Recall which is equal to the number of the actual class observations that were also predicted as positive divided by the total number of such observations in the actual class. It is important in scenarios where the consequences of a negative result are severe on the patient, as is often the case with HIV. On the Recall criteria, Logistic Regression, Random Forest, and SVM are highly accurate as seen by the values close to one. 0 meaning it is highly efficient in the screening of all the cases which link a particular disease. K-NN, has the lowest recall, thus meaning that k-NN will fail to retrieve many true positive cases, or records. Gradient Boosting and MLP keep the recall rates high, so it can be stated that these algorithms are rather accurate in identifying most of the true positive cases.

F1 Score reflects both precision and recall and is recommended when dealing with a situation of imbalance in the classes. Precision and recall measures are important for hyper-parameter detection because it is a balanced task where both metrics are significant, and the two models Logistic Regression and Random Forest have the highest F1 as a result. SVM also has good F1 score which further support the good results of this model. Again, k-NN has less F1 measure that was caused by less precise and recalled results. The F1 scores of Gradient Boosting and MLP are slightly lower than those of Logistic Regression and Random Forest, however, they seem to still be competitive. Like the above findings, Logistic Regression and Random Forest have always proven themselves discreet throughout the considered parameters.

SVM also performs well, mainly accuracy, precision and recall, meanwhile it also can be seen that the performance of using independent language features is better than using combined language features. K-NN is slower in all measures and seem not appropriate for this dataset or require hyper parameter-tuning. Gradient Boosting and MLP yield good accuracy, precision, recall, and F1 score values and hence can be considered as good substitutes for the Logistic Regression and Random Forest techniques. By use of this detailed analysis, one can know the various strengths and weaknesses of every model in order to choose the right one for use in classification of a given set of data. When choosing a model, certain of the parameters may be valued over others, depending upon the needs of a particular application, e.g. precision over recall or vice versa, and/or false positive rate over the false negative rate.

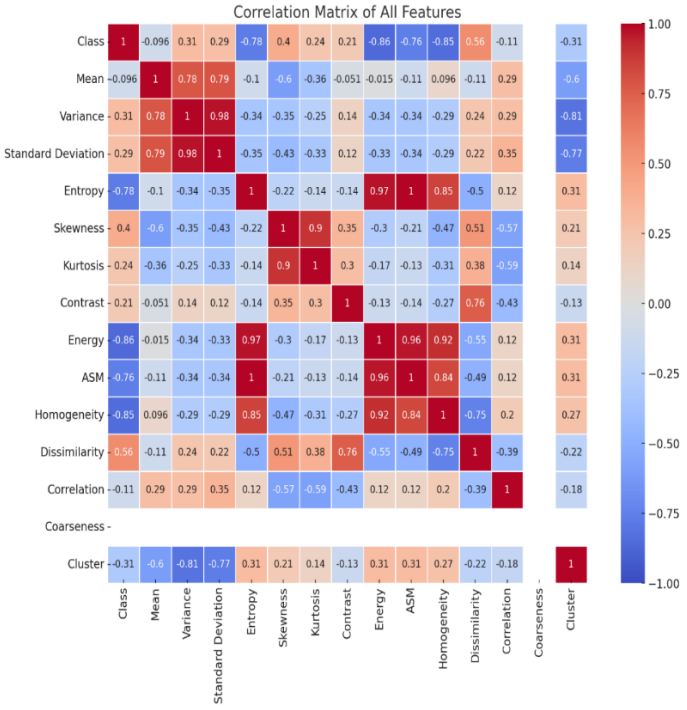

Fig.6. Correlation Coefficients of the Features in the Dataset

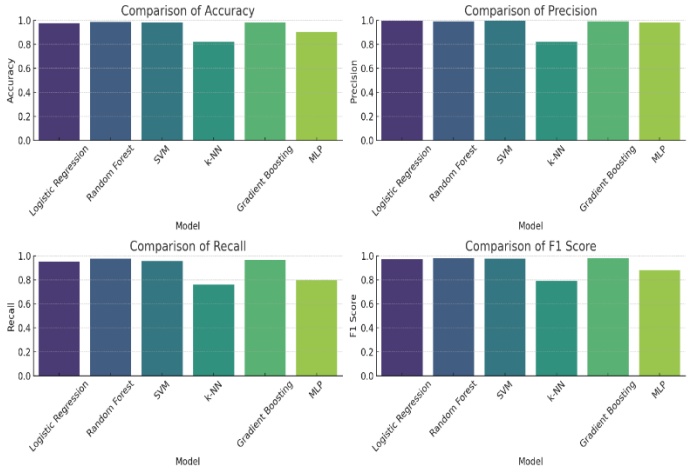

Fig.7. Comparative Analysis of MLP with Machine Learning Approaches

6. CONCLUSION

In this article, we have made and presented a detailed comparison on different machine learning models of our chosen ACR challenge for cancer detection with particular emphasis on the best-developed MLP model against some selected models. Therefore, the performance measures being assessed are accuracy, precision, recall, and F1 score, which give a comprehensive analysis of the models' performance. The main findings indicate that Logistic Regression and Random Forest are the most robust models that provide the greatest results in all indicators for cancer detection. SVM also performs fairly, especially in terms of precision and recall – both of which are significant in a disciplinary approach towards diagnosing the ailment. K-NN is again worse than all the others across the metrics; it might be incompatible with the given dataset or requires hyper parameter optimization. A high accuracy, precision, recall, and F1 score make Gradient Boosting and MLP reliable substitutes to Logistic Regression and Random Forest.

Explaining the details of the models helps to conclude which of them is stronger and which one is weaker and, thus, to choose properly when facing the problem of classification. The selection of model should allow the need for requirements of the application, whether it is more critical to recall more information when classifying or is more important to distinguish between the two classes with the minimum of false alarms or misses. Since early detection of cancer is crucial in treatment, the models that can sustain a good balance of both metrics are recommended to be used in the clinical setup, which encompasses the Logistic Regression, Random Forest, Gradient Boosting, and MLP. Also, when there are multiple layers in MLPs people often come across the problem of overfitting and this is very detrimental particularly when working with small or imbalanced medical datasets as may be observed in cancer detection. It negates the ability of the model to learn noise instead of developing patterns that can be used in any other environment. There is also a problem of data imbalance where benign cases are vastly more frequent than malignant ones; thus, to avert rendering overly optimistic results, resampling or class weighting has to be applied in a model to ascertain that the results obtained are not skewed. To overcome the problems the EDA analysis is conducted on dataset.

7. FUTURE SCOPE

This paper specifically compares the optimized MLP with the other machine learning methods applied for cancer detection with the incorporation of more enhanced deep learning methods like CNN and RNN. Applying ideas from transfer learning in pre-trained models may improve detection especially in situations of scarce data. More specifically, machine learning interpretable models will assist clinicians when they wish to comprehend the rationale behind model decisions. Models can be generalized to include other cancer types; additionally, more datasets can be added to the project. Finally, such methods as the real-time detection systems in clinics are crucial in early diagnosis and treatment planning [24].

REFERENCES

[1] A. Basha, and Niha Kamal, "Mask Region-based Convolutional Neural Network and VGG-16 Inspired Brain Tumor Segmentation", *Scientific Reports*, Vol. 14, pp. 1-9, 2024.

[2] H. Kokcam, Omer Mirac, Aytug Boyaci, and Muhammed Emre Colak, "Deep Learning based Brain Tumor Classification for MR Images using ResNet50", *Proceedings of International Symposium on Digital Forensics and Security*, pp. 1-6, 2024.

[3] A. Badshah and Shah Hussain, "An Efficient Classification Techniques for Brain Tumor Using Features Extraction and Statistic Methods with Machine Learning Algorithms", *Proceedings of International Symposium on Computer Science, Digital Economy and Intelligent Systems*, pp. 1-6, 2022.

[4] Rasool and Mohammed, "A Hybrid Deep Learning Model for Brain Tumour Classification", *Entropy*, Vol. 24, pp. 799-820, 2022.

[5] D.S. Vinod, "Ensemble Technique for Brain Tumour Patient Survival Prediction", *IEEE Access*, Vol. 12, pp. 19285-19298, 2024.

[6] Amran and Gehad Abdullah, "Brain Tumor Classification and Detection using Hybrid Deep Tumor Network", *Electronics*, Vol. 11, pp. 1-9, 2022.

[7] A. Krishna and Bandi, "A Brain Tumor Identification using Fully Convolution Neural Networks in the Deep Learning", *Proceedings of International Conference on Materials Science, Engineering and Chemistry*, Vol. 392, pp. 1-6, 2024.

[8] N.V.S.L. Narasimham, "Fuzzy Logic-Based System for Brain Tumour Detection and Classification", *Proceedings of International Conference on Image and Video Processing, Computer Vision and Pattern Recognition, Optimization and Control*, pp. 1-14, 2024.

[9] M. Indumathi and M. Saranya, "A Survey on Different Methods for Brain Tumor Detection", *Naturalista Campano*, Vol. 45, pp. 2967-2972, 2024.

[10] Marzieh Ghahramani and Nabiollah Shiri, "Brain Tumour Detection in Magnetic Resonance Imaging using Levenberg-Marquardt Backpropagation Neural Network", *IET Image Processing*, Vol. 17, No. 1, pp. 88-103, 2023.

[11] S. Ghosh, S. Mondal and B. Ghosh, "A Comparative Study of Breast Cancer Detection based on SVM and MLP BPN

Classifier", *Proceedings of International Conference on Automation, Control, Energy and Systems*, pp. 1-4, 2014.

[12] A. Das and Amita, "Detection of Liver Cancer using Modified Fuzzy Clustering and Decision Tree Classifier in CT Images", *Pattern Recognition and Image Analysis*, Vol. 29, pp. 201-211, 2019.

[13] Z.H. Choudhury and Ziaul Haque, "Integrating Random Forest, MLP and DBN in a Hybrid Ensemble Model for Accurate Breast Cancer Detection", *International Journal of Innovative Science and Research Technology*, Vol. 8, No. 7, pp.1-9, 2023.

[14] A. Rezaeipanah, "Design of Ensemble Classifier Model based on MLP Neural Network for Breast Cancer Diagnosis", *Inteligencia Artificial*, Vol. 24, No. 67, pp. 147-156, 2021.

[15] Jabeen Sultana and Abdul Khader Jilani, "Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers", *International Journal of Engineering and Technology*, Vol. 7, pp. 22-26, 2018.

[16] Xin Yu, Nazia Hameed and Jeremie Clos, "An Investigation of XGBoost-based Algorithm for Breast Cancer Classification", *Machine Learning with Applications*, Vol. 6, pp. 1-6, 2021.

[17] V. Sridevi, "Breast Cancer Examination in Digitized Mammograms using Integrated K-Means Clustering with Garbor Filter and Shrunk Kernel KNN Method", *Indian Journal of Science and Technology*, pp. 2444-2454, 2024.

[18] Nair, S. Sneha, V.N Meena Devi and Saju Bhasi, "Enhanced Lung Cancer Detection: Integrating Improved Random Walker Segmentation with Artificial Neural Network and Random Forest Classifier", *Heliyon*, pp. 1-9, 2024.

[19] A. Kumar and Tajinder, "Pre-Examination of Breast Cancer Dataset using Exploratory Data Analysis Approach", *Proceedings of International Conference on Computational Intelligence and Computing Applications*, Vol. 1, 2024.

[20] Vikram Verma, Alankrita Aggarwal and Tajinder Kumar, "Machine and Deep Learning Approaches for Brain Tumor Identification: Technologies, Applications, and Future Directions", *Proceedings of International Conference on Computational Intelligence and Sustainable Engineering Solutions*, pp. 1-6, 2023.

[21] Pankaj Kumar, and Karan Jain, "The Deep Learning Model for the Examination of Brain Tumor", *Proceedings of International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation*, Vol. 2, pp. 1-6, 2024.