International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Deep Learning and Multiclass Machine learning Classifier Approach for Predicting Primary Tumors

**Md Mehedi Hasan[1], Senjuti Rahman[2], Dr. Ajay Krishno Sarkar[3]**

[1]Electronics & Telecommunication Engineering, RUET,Rajshahi-6204
[2]Electrical & Electronic Engineering, Ahsanullah University of Science & Engineering, Dhaka
[3]Electrical & Electronic Engineering, RUET, Rajshahi-6204

## Abstract

Deep Learning (DL) and Machine Learning (ML) have the great prospect to play a significant role in the medical field in disease prediction. The tumor or cancer is one of the major health issues that each nation is currently dealing with, and it is the topic of this essay. The prediction of unidentified primary tumors in the dataset is delineated in this paper. Given that it provides significantly higher accuracy than binary classifiers, different multiclass classifier such as K-Nearest Neighbor (KNN), CatBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier, Light Gradient Boosting Machine, Ada Boost Classifier, Decision Tree Classifier, SVM - Linear Kernel, Naive Bayes and Deep neural networks (DNN1, DNN2, and DNN3) are used to categorize multiclass datasets available in the UCI machine learning repository. Among the stated machine learning classifiers, the k-Nearest Neighbor (KNN) had the highest classification accuracy of 92.92%. The three layer deep neural network (DNN2), among deep learning techniques, had produced the best accuracy of 97.66% using the chosen features as input. The gathered results from this work showed that deep neural networks outperformed machine learning techniques.

**Keywords:** Tumors, Classifiers, KNN, DNN, Performance Parameters.

## 1. Introduction

Cells are the relatively tiny building blocks that make up the body's tissues and organs. Cancer is such a disease that is the cause of the damage of the cells. Sometimes primary cancer, where it first began to grow, can spread to other parts of the body and cause new cancers (secondary cancers)[1]. The primary tumor is usually the one that can be removed easily. Finding it is crucial, because it might spread and cause secondary tumors, which are linked tumors. Even after tests are performed, doctors are unable to determine the location of the primary cancer when a secondary cancer is diagnosed. Most of the time, primary cancer is unidentified [2].

Several studies have been reported on the use of artificial neural networks (ANN) and machine learning techniques for survivability analysis and forecasting. The healthcare industry makes extensive use of data mining techniques like clustering, classification, regression, association rule mining, and CART (Classification and Regression Tree). M. A. Khaleel et al. [3] primarily focused on the analysis of data

mining techniques needed for medical data mining, particularly to identify locally prevalent diseases like heart conditions, lung cancer, breast cancer, and so forth. Additionally, it shaded light on the significance of regionally prevalent patterns and the mining methods employed for that purpose. According to M. Khan's research [4], one of the terrible diseases that claims the majority of lives worldwide is cancer. Blood cancer symptoms and staging had been studied in that work. Using the linear regression algorithm, the areas having the highest cancer rates was determined, also the ratio of cancer cases involving men versus women, and whether factors such as lifestyle, diet, education, marital status, and place of residence play a significant role in the cancer pattern was studied. The work reported in [5] stated, the clustering model is better suited for pattern recognition if the dataset contains unlabelled classes or features, meaning that, the performance of each model varies depending upon the type of dataset used. S. Vijiyarani et al.'s [6] primary focus was on the prognosis of breast cancer, diabetes, and heart disease. The heart disease dataset was examined using the Naive Bayes, K-NN, and Decision List algorithms. In order to train the model from various angles and perspectives to achieve good results and scores, Sobhaninia et al. [7] used a LinkNet network with a CNN model for segmentation using brain MRI. Although they achieved an accuracy of 79%, the model is complex. A method for classifying and detecting brain tumors based on CNN was created by Sajjad et al. [8]. The accuracy achieved by the authors, who used a Cascade CNN algorithm for segmenting brain tumors and a fine-tuned VGG19 algorithm for classifying tumors, was 94.58%. In order to address the issue of overfitting, Kumar et al. [9] proposed a brain tumor method using the ResNet50 CNN model and global average pooling, and achieved an average accuracy rate of 97.48%.

The right categorization model should be used when diagnosing a primary tumor. The primary tumor dataset has been examined using different multi class algorithms to look for patterns that had not previously been noticed in this paper. This study is to investigate the various deep learning (DL) and machine learning (ML) approaches used for tumor classification. Nine different classification algorithms were used in this work to identify the primary stage tumor, including Naive Bayes, KNN, CatBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier, Light Gradient Boosting Machine, Ada Boost Classifier, Decision Tree Classifier, SVM - Linear Kernel, Naive Bayes, and Neural Networks(DNN1, DNN2 and DNN3). To demonstrate the novelty of this work, the performance metrics of the most accurate approaches were compared with those of the earlier works.

## 2. Description of the Dataset

Data collection is the first stage of system processing, and to measure the performance parameters of the two proposed approaches,UCI ML primary tumor dataset was utilized in this research.There are 339 instances, 18 attributes, and one class attribute in the dataset. There are 22 different classes of primary tumors in total. Furthermore, the remaining characteristics show the origins of primary tumors while the "goal" field refers to the patient's having primary stage tumor or not. The description of the main attributes of the features is given in Table I. Dataset attributes are characteristics that are used for systems. Before moving on to the analysis stage, data was preprocessed in this work. The handling of missing value is one of them. We must transform some categorized values using dummy value means in the form of "0" and "1" for our work. A balanced set of data is necessary for accurate results. By data balancing, we have equalized both target classes, improving the accuracy of the validation.

Table1. Description of the Attributes of the Dataset

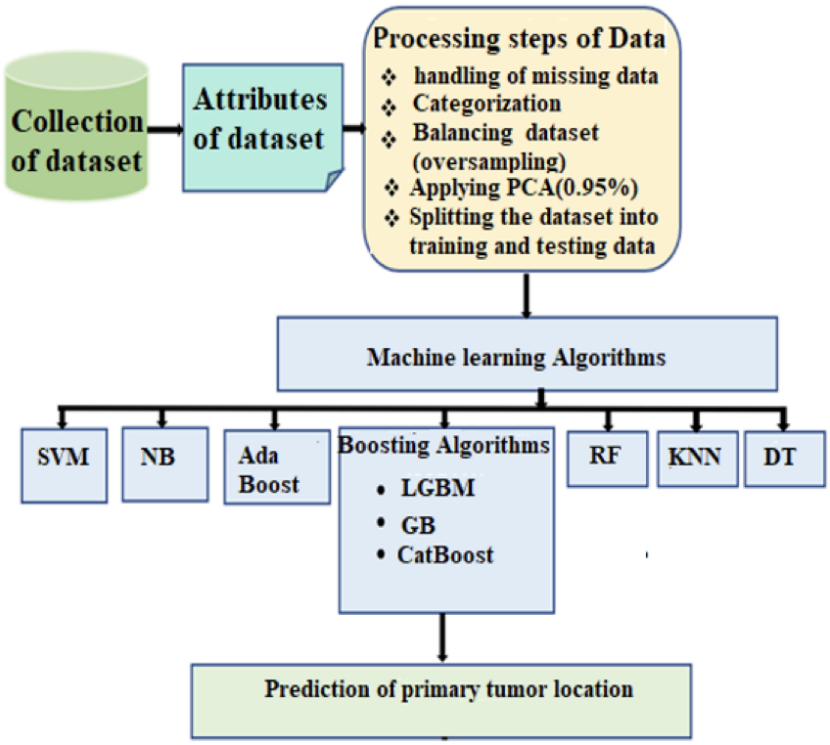
| S.No. | Attribute | Description of the Parameters |
|---|---|---|
| 1 | Age | Demographic data related to patient's age (<30, 30-59, >=60) in years |
| 2 | Sex | Demographic data related to the Gender of the patient<br>Male-1,<br>Female-0 |
| 3 | histologic-type | Multivariate data of 3 types<br>type1:epidermoid, type 2: adeno, type 3: anaplastic |
| 4 | degree-of-differ | Multivariate data of 3 degrees<br>well, fairly, poorly |
| 5 | Bone | Yes-1(tumor detected), No-0(not detected) |
| 6 | bone-marrow | Yes-1(tumor detected), No-0(not detected) |
| 7 | Lung | Yes-1(tumor detected), No-0(not detected) |
| 8 | Pleura | Yes-1(tumor detected), No-0(not detected) |
| 9 | peritoneum | Yes-1(tumor detected), No-0(not detected) |
| 10 | Liver | Yes-1(tumor detected), No-0(not detected) |
| 11 | Brain | Yes-1(tumor detected), No-0(not detected) |
| 12 | Skin | Yes-1(tumor detected), No-0(not detected) |
| 13 | Neck | Yes-1(tumor detected), No-0(not detected) |
| 14 | supraclavicular | Yes-1(tumor detected), No-0(not detected) |
| 15 | axillar | Yes-1(tumor detected), No-0(not detected) |
| 16 | mediastinum | Yes-1(tumor detected), No-0(not detected) |
| 17 | abdominal | Yes-1(tumor detected), No-0(not detected) |
| 18 | Class | Location of tumor (lung, head & neck, esophasus, thyroid, stomach, duoden&sm.int,colon, rectum, anus, salivary glands, pancreas, gallblader, liver, kidney, bladder, testis, prostate, ovary, corpus uteri, cervix uteri, vagina, breast) |

## 3. Approaches for Classification of Primary Tumor

### 3.1 Approach I: Finding the Best Machine Learning Algorthim for Detecting Tumor

Nine machine learning algorithms were used in this study, including K-Nearest Neighbor (KNN), CatBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier, Light Gradient Boosting Machine, Ada Boost Classifier, Decision Tree Classifier, SVM-Linear Kernel, Naive Bayes and Deep neural networks (DNN1, DNN2, and DNN3). The flow chart of the proposed work is shown in Figure 1. Performance metrics (Accuracy, Area under the curve (AUC)) were recorded and a comparison chart is shown in Figure 2.

Figure 1. Flow Chart of the Steps of the Proposed Work for Implementing Machine Learning Algorithms.




### 3.1.1K-Nearest Neighbor

Astraightforward supervised learning algorithm called K-Nearest Neighbor [10] is used to model classifications and regressions. It centres its classification on the fundamental presumption that, data points of the same kind are located close to one another. The closest K neighbors for any given data point are encircled in a perimeter, and the data sample is classified according to the class of the maximum number of neighbors. K stands for the number of neighbors chosen to take the votes from. In this study, We achieved the highest accuracy by using KNN algorithm.The name of the parameters and default value for this multiclass classifier is described below in table II. In our study, k = 5 was utilized.

Table2. Description of the Parameters of KNN Classifier

| **Name of the Parameter** | **Utilized Values** |
|---|---|
| Leaf_size | 30 |
| metric | minkowski |
| metric_params | none |
| n_jobs | -1 |
| n_neighbors | 5 |
| p | 2 |
| weights | uniform |

### 3.1.2Naïve Bayes

The Naive Bayes probabilistic classifier is built on the Bayesian theorem (NB). The classifier is referred to as naive because it operates under the strong features independence assumption. The key distinction between the various versions of NB found in the literature [11] is, how the likelihood of the intended class is computed. Simple Naive Bayes, Gaussian Naive Bayes (which was used in this study), Multinomial Naive Bayes, Bernoulli Naive Bayes, and Multi-variant Poisson Naive Bayes are some of these variations.

### 3.1.3The Support Vector Machine

One of the most popular machine learning algorithms, Support Vector Machine (SVM) [12], offers adequate accuracy while requiring less processing power. This algorithm discovers a hyper plane to categorize data points in an N-dimensional space with the greatest possible margin, or the distance between data points of the various classes [13]. The classification's decision boundary is the hyperplane. SVM aims to categorize the data by creating a function that divides the data points into their corresponding labels with the least amount of error and greatest (maximum) possible margin.

### 3.1.4 Light Gradient Boosting Machine, CatBoost, Gradient Boosting

CatBoost is an algorithm for decision trees that uses gradient boosting. LightGBM has a number of benefits, including better accuracy, lower memory usage, higher efficiency, and faster training speeds. All three of the algorithms under consideration employ gradient boosting techniques (CatBoost, Gradient Boosting, and LightGBM). CatBoost generates symmetric (balanced) trees as opposed to XGBoost and LightGBM.

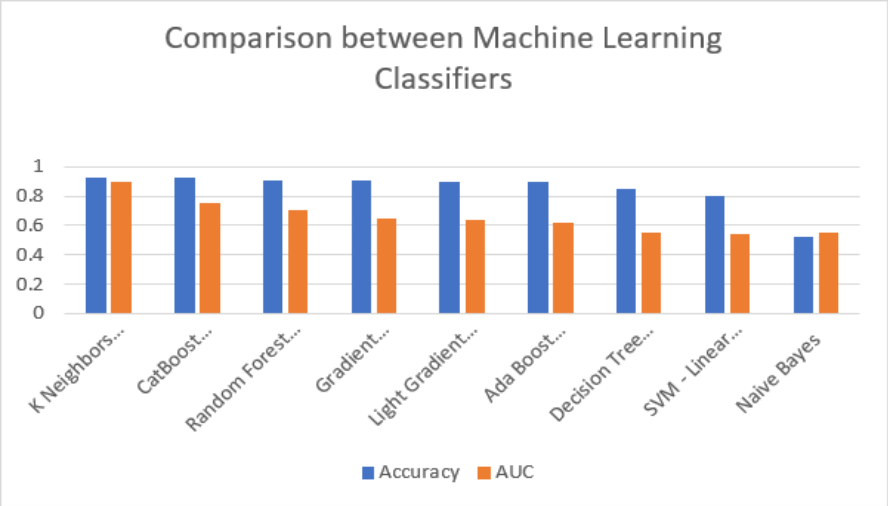
### 3.1.5Ada Boost

Machine learning ensemble methods that use the boosting technique are known as the AdaBoost algorithms, also referred to as Adaptive Boosting. Since the weights are redistributed to each instance, instances that were incorrectly classified are given higher weights, hence the term "adaptive boosting." The 100 DT(Decision Tree) algorithm was combined with the CART (Classification and Regression Tree) algorithm. Each time the boosting procedure was repeated, each classifier was given a weight of 1.

### 3.1.6Random Forest and Decision Tree

Bagging, also referred to as Bootstrap Aggregation, is the ensemble method used by random forest. The Random Forests (RF) classifier is a type of ensemble approach that combines numerous decision tree predictions. RF creates trees at random by selecting attributes at each node. The output of the ensemble are tree votes for the most popular class. The random forest method is more robust to errors and outliers. As a result, the overfitting problem is not present while DT has an overfitting issue. Another well-liked and simple to use supervised learning algorithm is the Decision Tree (DT) [14]. It resembles a tree diagram in that it has branches, leaf nodes, and internal nodes. The independent variables in a classification or regression problem are represented by the internal nodes. The solution to the issue, or the dependent variable, is represented by the leaf nodes [15].

Figure2. Comparison Between the Machine Learning Algorithms.




## 3.2 Approach II: Deep neural networks (DNN1, DNN2, DNN3) for Classifying Primary Tumor

Deep learning is a branch of machine learning that makes use of artificial neural networks. Examples of deep learning architectures include convolutional neural networks, deep belief networks, recurrent neural networks, and deep neural networks (DNN). These are widely used in many different research areas, such as speech recognition, natural language processing, audio recognition, computer vision, gaming, and many others[16]. A DNN is made up of an input layer, several hidden layers, and an output layer [17]. Backpropagation is used to train the network and reduce the difference between the desired and actual output. Three DNN models have been put into practice for the analysis. The three models have three (DNN1), four (DNN2), and six (DNN3) hidden layers, respectively, with a sigmoid function at the output.Figure 3 displayed the proposed deep neural network models. Figure 4 compares the evaluation criteria, including recall, accuracy, and precision for the three DNN models. It is clear from the figure that DNN2 performs better than DNN1 and DNN3.For the approaches, we utilized 'adam' Optimizer, 'binary_crossentropy' as loss function, 'accuracy' as metrics, and Epochs =100 to achieve the desired results.

# 4. Evaluation of Metrics and Results

## 4.1 Evaluation Metrics

In this study, the performance and efficacy of the classifiers were evaluated using five statistical measures, namely, accuracy, precision, recall/sensitivity (recall and sensitivity are the same in binary classification), f-1 score, and AUC curve. The corresponding equations of those measures are provided as follows.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Figure 3. Three approaches of applying DNN model; (a) DNN1, (b) DNN2, and (c) DNN3.

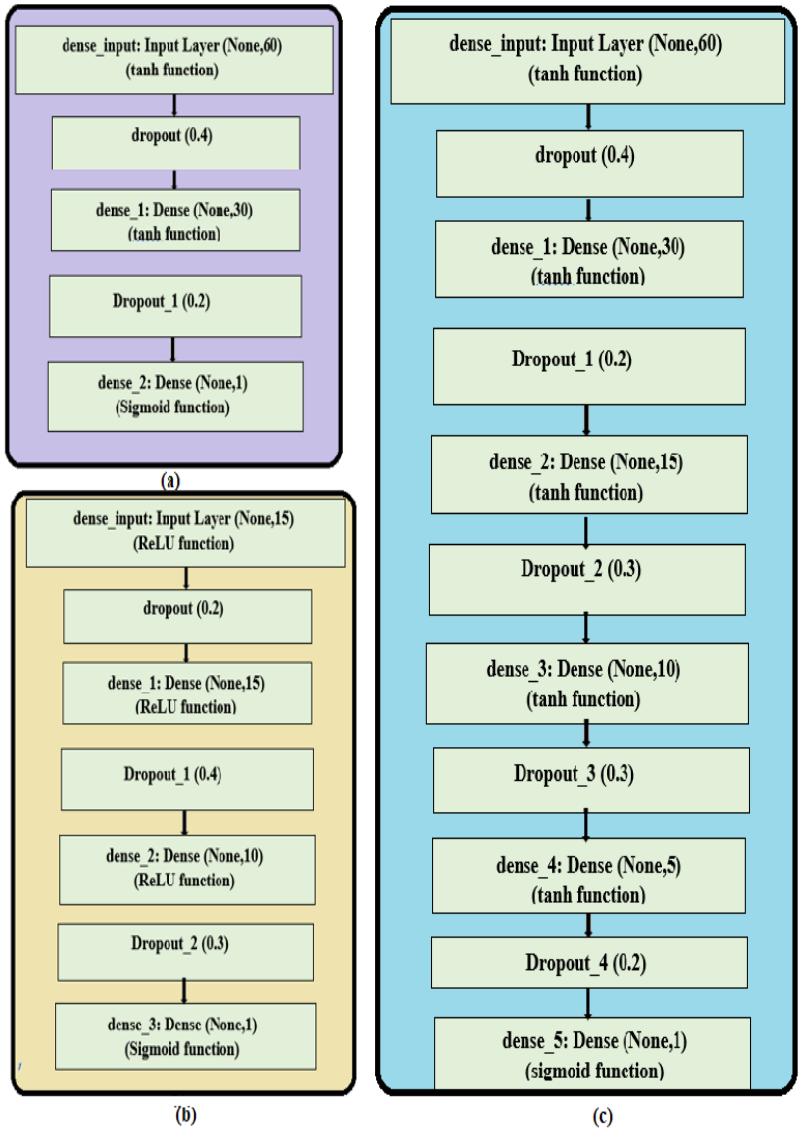
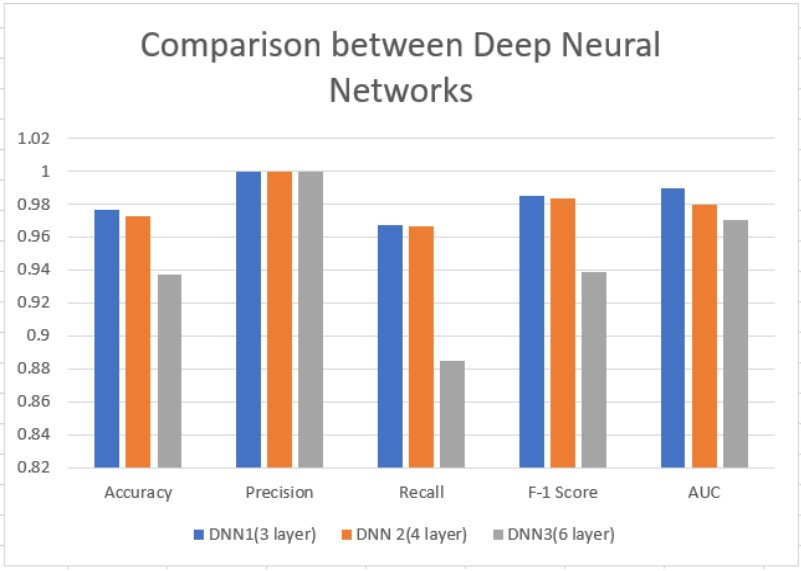



Figure 4. Comparison Between Deep Neural Networks.



$$\text{Recall/Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1 Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}$$

Where, TP = true positive
FN = false negative,
FP = false positive
TN =true negative
This study used AUC curves to examine how well the probabilities from the positive classes and the negative classes could be separated from one another. The degree of True Positive and False Positive rate is usually represented by the AUC curve, which also shows how well the model performed overall.

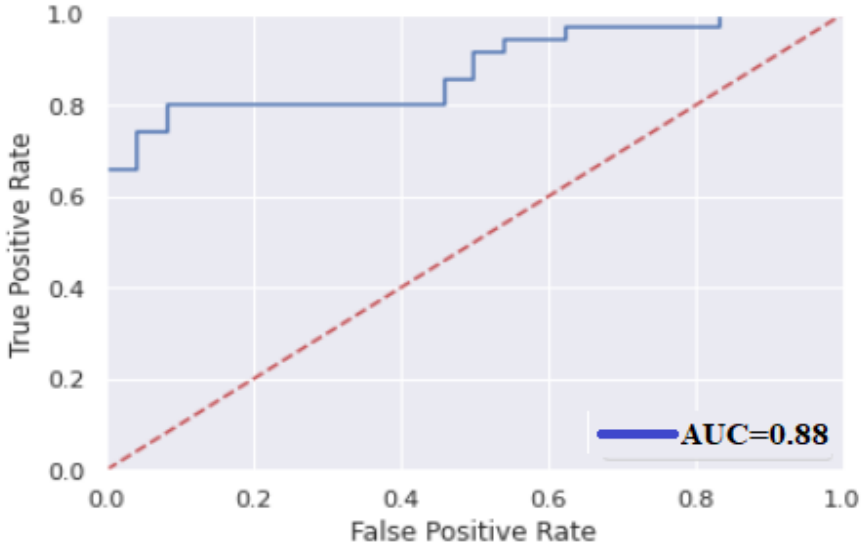
## 4.2 Results

KNN outperforms other classifiers in terms of accuracy (92.92%) among machine learning algorithms, whereas DNN2 demonstrated promising results (accuracy of 97.66%) among deep learning techniques. The AUC curve of the two best performed method is shown in Figure 5. From the comparison, it is clear that DNN2 algorithm outperforms KNN algorithm in all aspects.

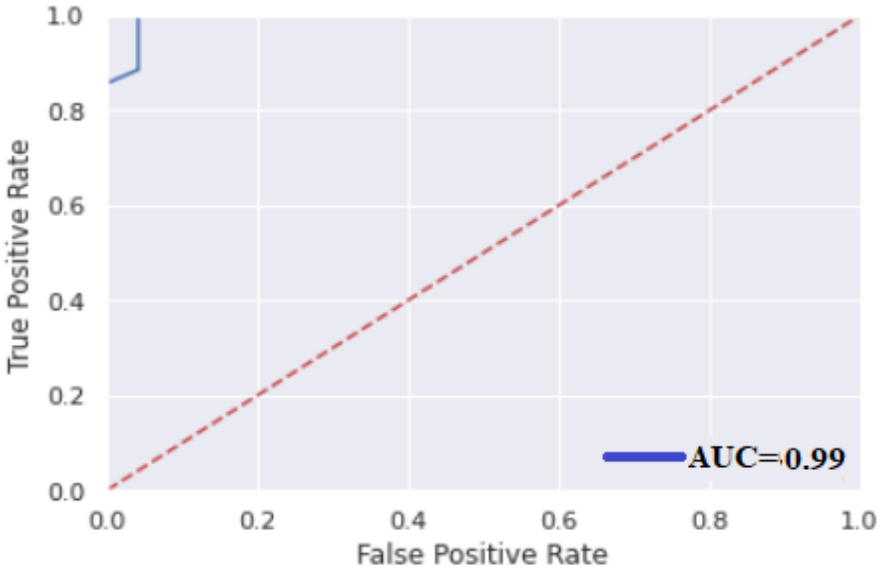
## 5 Comparison with the Existing Work

It becomes clear that the proposed method for identifying primary stage of tumor is novel when it is contrasted with earlier studies in the area. The comparison is shown in Table III.

Figure 5. The AUC Curve of the Two Best Performed Methods; (a) KNN classifier, (b) DNN2.




(a)




(b)

Table III. Comparison of the Proposed Methods with the Related Work

| **Related Work** | **Objectives** | **Source of Data** | **Machine Learning/Deep Learning Approaches** | **Outcomes** |
|---|---|---|---|---|
| [1] | Prediction of Primary tumor | UCI machine learning repository | Multiclass random forest, naïve bayes | Random forest accuracy= 85.70%, Naïve bayes accuracy =54.23% |
| [7] | Detection of Brain tumor | - | LinkNet network with a CNN model | Accuracy= 79% |
| [18] | Classification of brain tumor | - | deep CNN models with SVM and KNN | Accuracy =97.67% |
| [19] | Detection of Brain tumor | - | Gray Level Co-occurrence and SVM | Accuracy = 94.52% |
| **Proposed work (With KNN)** | Prediction of Primary tumor | UCI machine learning repository | Ada Boost, LightGBM CatBoost, Gradient Boosting, RF, DT, KNN, SVM, NB | **KNN Accuracy= 92.92%** |
| **Proposed work (With DNN2)** | Prediction of Primary tumor | UCI machine learning repository | DNN1, DNN2 and DNN3 | **DNN2 Accuracy: 97.66%** Precision= 1.0, Recall = 0.9672 |

## 6 Conclusion

A patient's quality of life can be improved by an early diagnosis of tumor which is crucial for clinical management and cancer surveillance and a significant topic of research today. Using machine learning and Deep Learning techniques, the main emphasis of this work was onto primary tumor prediction using various algorithms and a combination of several targets' attributes. Depending on the dataset, a specific ML and DL algorithm was chosen, and it was typically discovered that binary classifiers were accurate with no more than two classes, whereas multiclass classifiers with binary classifiers, like KNN, were more accurate with more classes. Additionally, it has been discovered that the deep neural networks (specially DNN2) technique increases the accuracy of the chosen classifier's results. The work can be expanded and improved in the future to automate the prediction of primary tumors. The parameters of the learning algorithms weren't tuned to produce the greatest outcomes with our data, which is a good concluding point that we would like to make. So, there's still room for development.

## 7 Authors' Biography

**MD MEHEDI HASAN** received the B.Sc. degree from Rajshahi University of Engineering and Technology, Bangladesh, in 2016 in electronics and telecommunication engineering. From 2017 to 2019, he was working as a Junior Software Engineer at Smart Aspects Ltd. In Dhaka, Bangladesh. Recently, from December, 2019 he joined as a Software Engineer at Zantrik, Dhaka, Bangladesh. His research interests include Biomedical Engineering, Machine Learning, and Deep Learning Biomedical Engineering, Machine Learning, Deep Learning, Computer Vision, Data Science. His working area covers Android Mobile Application Development, Web Development, Database Management, Machine Learning, Deep Learning. He has a good grasp of some programming languages, such as Java, C#, Python, C++, SQL. He has completed two online courses, AI for Medical Diagnosis- an online non-credit course authorized by Deep Learning.AI and offered through Coursera and Machine Learning Projects for Healthcare on Udemy. He has 5 international conference papers in ICEEE 2017, 4IREF 2022, ICECE 2022, ICCIT 2022 and two journals have been accepted at the European Journal of Electrical Engineering and Computer Science (EJECE) and onejournal has been accepted at International Research Journal of Engineering and Technology (IRJET, 2022). Three of his recent works are under process.

**SENJUTI RAHMAN** received the B.Sc. degree from Rajshahi University of Engineering and Technology, Bangladesh, in 2016 in electronics and telecommunication engineering. Currently, she is pursuing the M.Sc. degree in electrical and electronic engineering from the same university. From 2018 to 2022, she was working as a lecturer in the department of EEE in Eastern University (EU), Dhaka, Bangladesh. From 25th July, 2022 she joined Ahsanullah University of Engineering and Technology as a Lecturer in the EEE department. She has a total of 7 conference papers in ICECTE 2016, ICEEE 2017, ICRPSET 2022, ICECE 2022, ICCIT 2022, 4IREF 2022 and two journals have been accepted at the European Journal of Electrical Engineering and Computer Science (EJECE)and onejournal has been accepted at International Research Journal of Engineering and Technology (IRJET, 2022). Three of his recent works are under process. Her research interests include Biomedical Engineering, Machine Learning, and Deep Learning.

**DR. AJAY KRISHNO SARKAR** has received Ph. D in Electronic and Computer Engineering from Griffith University, Australia, M. Sc in EEE from Japan and B. Sc. in EEE from RUET, Bangladesh. He is currently working as a professor in the department of Electrical and Electronic Engineering (EEE) at Rajshahi University of Engineering and Technology (RUET), Rajshahi-6204, Bangladesh. He is currently the member of IEEE; Institute of Engineers Bangladesh (IEB) and he had the position in different organizing and technical committees in different international conferences at Bangladesh and abroad. He is a reviewer of several journals such IEEE Access, IEEE Photonics Journal, Computers and Electronics in Agriculture etc. and technical papers submitted in different international conferences in Bangladesh and abroad. His research interests include Sports and Biomedical Engineering, Photonic Crystal Fiber and Biosensors, Microwave and RF circuits & Devices, Microwave absorptions, Thin Films.

## 8 References

1. M. Naib and A. Chhabra, "Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining," *International journal of computer applications*, vol. 96, no. 8, p. 9, Jun. 2014.
2. N. Bhuvaneswari and S. Yamuna, "Information extraction of predicting blood cancer", *International Journal of Computer Science*, Vol. 1, Issue 4, 2013.
3. M. Khaleel, S. Pradham and G.N. Dash, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, Issue 8, 2013.
4. M. Khan, S. Qamar and L. Massin, " A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining", International Journal of Applied Engineering Research, Vol.7, issue.11, 2012.
5. K. Lokanayaki and A. Malathi, "Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis", International Journal of Computer Applications, Vol. 77, No.5, 2013.
6. S. Vijiyarani and S. Sudha ,"Disease Prediction in Data Mining Technique – A Survey", *International Journal of Computer Applications & Information Technology*, Vol. II, Issue I, pp. 2278-7720, 2013.
7. Z. Sobhaninia et al., "Brain Tumor Segmentation Using Deep Learning by Type Specific Sorting of Images.," *arXiv: Computer Vision and Pattern Recognition*, Sep. 2018.
8. M. Sajjad, S. Khan, S. W. Baik, W. Wu, and A. Ullah, "Multi-grade brain tumor classification using deep CNN with extensive data augmentation," *Journal of Computational Science*, vol. 30, pp. 174, Jan. 2019.
9. R. L. Kumar, J. Kakarla, B. V. Isunuri, and M. Singh, "Multi-class brain tumor classification using residual network and global average pooling," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13429, Jan. 2021.
10. Md. A. B. Siddique, S. Sakib, and Md. A. Rahman, "Performance Analysis of Deep Autoencoder and NCA Dimensionality Reduction Techniques with KNN, ENN and SVM Classifiers," *arXiv: Learning*, Dec. 2019.
11. T. Wasif, M. I. U. Hossain and A. Mahmud, "Parkinson disease prediction using feature selection technique in machine learning," *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-5, 2021.

12. Sakib, M. A. B. Siddique, and M. A. Rahman, "Performance Evaluation of t-SNE and MDS Dimensionality Reduction Techniques with KNN, ENN and SVM Classifiers," *IEEE Region 10 Symposium (TENSYMP)*, pp. 5–8, 2020.
13. K. M. Hasib et al., "A survey of methods for managing the classification and solution of data imbalance problem," *arXiv Prepr.*, 2020.
14. H. Cho et al., "A similarity study of content-based image retrieval system for breast cancer using decision tree," *Med. Phys.*, vol. 40, no. 1, pp. 12901, 2013.
15. K. M. Hasib, M. I. H. Showrov, and A. Das, "Accidental prone area detection in bangladesh using machine learning model,"*3rd International Conference on Computer and Informatics Engineering (IC2IE)*, pp. 58–62, 2020.
16. I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN computer science*, vol. 2, no. 6, Jan. 2021.
17. Aiello, S., Cliff, C., Roark, H., Rehak, L., Stetsenko, P., and Bartz, A., "Machine Learning with Python and H2O", *H2O. ai Inc*,2017.
18. H. Kibriya, R. Amin, A. H. Alshehri, M. Masood, S. S. Alshamrani, and A. Alshehri, "A Novel and Effective Brain Tumor Classification Model Using Deep Feature Fusion and Famous Machine Learning Classifiers," *Computational Intelligence and Neuroscience*, vol. 2022, p. 1, Mar. 2022.
19. P. Shanthakumar and P. G. Kumar, "Computer aided brain tumor detection system using watershed segmentation techniques," *International Journal of Imaging Systems and Technology*, vol. 25, no. 4, p. 297, Dec. 2015.