International Journal of Advanced Research in Engineering Technology & Science
ISSN: 2349-2819
Impact Factor: 7.10
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Email: editor@ijarets.org Volume-10 Issue-12 December-2023 www.ijarets.org

A Comparative Study of Supervised Machine Learning Algorithms: A Case Study on Osteosarcoma

Divya Chauhan
Rajiv Gandhi Government College,
Chaura Maidan, Shima, HP
India

Abstract: Cancer is one of leading causes of mortality across the world. Osteosarcoma is the most common type of primary bone cancer that usually develops in the long bones such as the tibia, femur or humerus. It mostly affects children, teenagers and young adults. It is originated in the osteoblast cells that form new bones. Increased use of modern technology like artificial intelligence, machine, and deep learning techniques create more accurate and trustworthy models for predicting and detecting osteosarcoma. The study evaluates five supervised machine learning algorithms like SVM, Naïve Bayes, k nearest neighbor classifier (kNN), Random Forest and Logistic Regression for multiclass tumor classification. Performance metrics including precision, Recall, F1- Score, accuracy and misclassified counts are analyzed across four classes: non-tumour, non-viable tumour, viable tumour, and a combined viable/non-viable tumour class.

Keywords: *Machine learning, classifiers, cancer,feature extraction*

I. INTRODUCTION

Machine learning is a field of artificial intelligence focused on developing algorithms and statistical models that enables computer system to learn patterns from the data and make prediction based on that given data without explicit rule based program. The mean goal of machine learning algorithms is to perform well on unseen or test data not just training data. Machine learning [1] has grown rapidly in recent years in the context of data analysis and computing that typically allows the applications to function in an intelligent manner. The integration of machine learning and cancer diagnosis holds the promise of improving patient outcomes, refining treatments and support well informed decision making.

Classification is a type of supervised machine learning where the goal is to predict a category for class label on basis of the training provided to the model. Common classification algorithms include k nearest neighbours (kNN), support vector machine (SVM), logistic regression, Gaussian Naive bayes, and Random Forest. Classification algorithms are widely used in medical image processing to detect disease from various scans. There are certain performance evaluation criteria for these algorithms. Moreover, the Machine Learning models are evaluated [2] based on certain Metrics parameters like Accuracy, Precision, Recall, F1-Score.

Osteosarcoma is an aggressive cancer of the bone that originates from immature bone-forming cells. It most often appears in the long bones of the legs and arms, particularly around the knee, where bone growth is fastest. The disease is seen most commonly in children, adolescents, and young adults, usually during periods of rapid growth. Machine learning enables early and accurate cancer detection by identifying subtle patterns in medical data that may not be visible to the human eye [3][4]. It improves diagnostic prediction of osteosarcoma by analyzing large datasets, reducing human error, and supporting decision making.

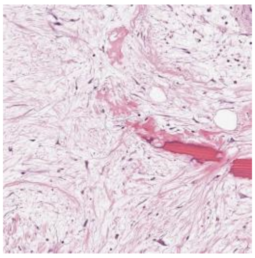
II. Literature Review

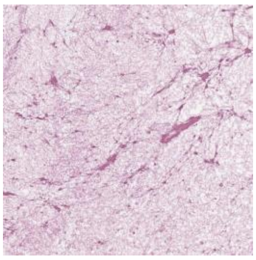
Ahmed M. A. et al [5] investigated the use of Machine Learning techniques to predict cancer risk by integrating genetic and lifestyle factors. The study utilized a structured dataset of 1,200 patient records and implemented a complete ML pipeline workflow, encompassing data exploration, preprocessing, feature scaling, model training, and evaluation. Nine supervised learning algorithms were evaluated and compared, including Logistic Regression (LR), Support Vector Machines (SVMs), Decision Tree (DT), Random Forest (RF), and ensemble methods. Among these, Categorical Boosting (CatBoost) achieved the highest predictive performance, with a test accuracy of 98.75% and an F1-score of 0.9820, outperformed others. Mohammad H.S. et al [6] examined the potential of Deep learning, particularly Convolutional Neural Networks, to improve ovarian cancer diagnosis by accelerating the diagnostic process and enhancing precision. The paper concluded that by addressing such research areas, further advancements can be made in leveraging other deep learning techniques to enhance the accuracy, efficiency, and reliability of ovarian cancer diagnosis. F. J. Shaikh and D. S. Rao [7] reviewed various machine learning and deep learning approaches applied to cancer prognosis. They highlighted differences in model training, prediction endpoints, cancer types under study, and overall performance across existing prediction models. Seeta Devi et al. [8] demonstrated that both traditional and modern machine learning algorithms can effectively predict breast cancer, with deep learning models such as Multilayer Perceptrons (MLP) showing superior accuracy. The study recommends adopting AI-based systems in clinical settings to support early diagnosis and improve therapeutic outcomes for breast cancer patients. K. N. Mishra et al. [9] underscored the link between cellular characteristics and cancer risk, stressing the significance of accurate data interpretation. Their experiments were conducted using the LUNA16 Lung Cancer dataset obtained from Kaggle. Kourou K. et al. [10] summarized the recent advancements in AI and ML applications within cancer research and medical oncology. The authors also discussed existing challenges and essential considerations that must be addressed before data-driven models can be effectively integrated into clinical practice. Pereira H. M. et al. [11] developed ML-based CT radiomic features to predict the onset of metastasis in osteosarcoma patients. The study included 81 participants, and the optimal classifier was selected and tested on unseen data to ensure an unbiased assessment of the final model's performance. Varlamova [12] discussed the prospects of machine learning technologies in medicine, reviewing methods such as CNN, decision trees etc. The study highlighted their applications in cancer detection, treatment planning, and survival prediction.

III. Results and Findings

Dataset Used

The study has retrieved around 20 GB of image data is from TCIA for cancer imaging archive. TCIA is a platform that hosts a comprehensive archive of cancer-related medical images, available for global access and download. The repository includes CT, MRI, and nuclear medicine scans. The primary file format used is DICOM (Digital Imaging and Communications in Medicine), an international standard for storing, transmitting, and processing medical imaging data. Figure 1 below illustrates sample osteosarcoma image tiles from the dataset.





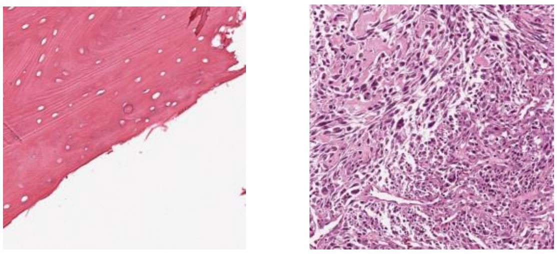


Figure1: Images of osteosarcoma used for the study

Preprocessing of data

In order to improve the data and to enhance the performance of machine leaning models, the data which is being fed to the classifier needs to be preprocessed. Data needs to be transformed into the format algorithms can understand and analyse effectively. The dataset used in the study is in .dcm file format which is downloaded with the help of NBIA data retriever. Further .dcm files are fed PearlMountain Image Converter to get .jp, .png and .tiff file format. Now the image dataset is transformed into smaller new manageable set of new features using feature extraction techniques. It uses many mathematical and statistical methods, like extracting texture or edges instead of raw pixel values, to combine original variable into new low dimensional values and remove noises and redundancy. The study uses python for transformation and processing. Pandas library has ample of techniques and algorithms to makes the process of extraction, filter and transformation of data quick and efficient. Dataset is rescaled to the range of 0 and 1 with the help of Scikit learn library of python.
After feature extraction of images in the dataset, there are four classes of tumor in the dataset Non Tumor, Non-viable Tumor, Viable , Viable: Non-viable.

Non tumour is a class that are not malignant; they may increase in size but lack the ability to invade adjacent structures or metastasize to distant sites.

Non Viable tumor denotes necrotic or biologically inactive tumour tissue that has lost its capacity for cellular growth and proliferation.

Viable tumor is characterized by the presence of intact epithelial tumour cells within the lymph node, indicating active cellular metabolism, proliferation, and potential for further progression.

Viable:non viable tumor is the class of tumor where it is not clear to which category the tumor belongs to. Whether it is Viable tumor or Non Viable tumor.

The percentage of number of occurrences of these classes in the dataset is shown in the figure 2 below in the form of pie chart.

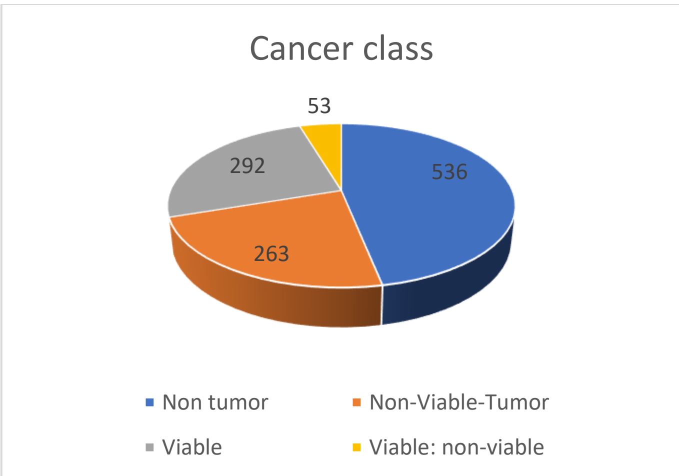


Figure 2: Occurrences of tumor cases in dataset

Classification of data

The dataset is separated into training and testing subsets. A stratified random split is applied to ensure that the class distribution is preserved across both sets, preventing all samples of a particular class from being placed into only one subset. Figure 3 below presents the Python code used in this study to perform the stratified random split.

```
split=StratifiedShuffleSplit(n_splits=1,test_size=0.2,random_state=50)
for train_index,test_index in split.split(X,y):
    X_train_set=X.loc[train_index]
    X_test_set=X.loc[test_index]
    y_train_set=y.loc[train_index]
    y_test_set=y.loc[test_index]
```

Figure 3: Stratified Shuffle Split

The data also had many missing values. So further transformation has been used to fed them to classfiers as many do not handle missing values. Pipelining allows to sequentially apply a list of transformers to preprocess the data and conclude the sequence with final predictor for predictive analysis. Column transformer is used in the paper to transform columns of an array or pandas DataFrame and the features produced by each transformer are then concatenated to create a unified feature space. Univariate imputer is used for completing missing values with simple strategies like mean and most frequent. MinMaxScaling scales and transforms each feature individually such that it falls within the given range on the training set, such as between 0 and 1. Finally the data is ready to be fed to the classifiers for the training purpose. Figure 4 below shows the chart of transformations used for one of the classifiers i.e. SVM.

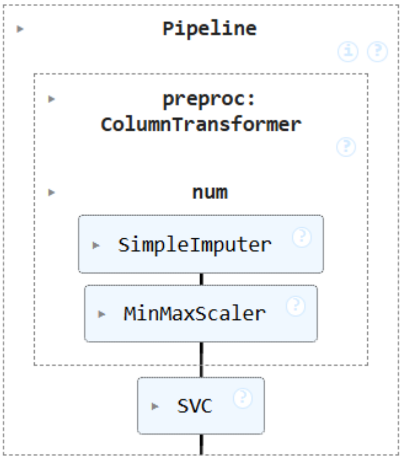


Figure 4: Transformation on dataset prior training the model

Results and discussion:

After the evaluation of five supervised machine learning algorithm for multi class tumour classification following results have been found. Table 1 below shows the evaluation matrics of 5 classifiers for tumour classification..

Table 1 Evaluation matric of classifiers

Class	Precision	Recall	F1 score	Support
		SVM		
Non tumour	0.88	0.91	0.89	107
Non Viable tumor	0.87	0.89	0.88	53
Viable tumor	0.78	0.88	0.83	58
Viable:non viable tumor	0.00	0.00	0.00	11
		kNN		
Non tumour	0.90	0.92	0.91	107

Non Viable tumor	0.86	0.91	0.88	53
Viable tumor	0.79	0.76	0.77	58
Viable:non viable tumor	0.25	0.18	0.21	11
Random forest				
Non tumour	0.88	0.91	0.89	107
Non Viable tumor	0.87	0.89	0.88	53
Viable tumor	0.78	0.88	0.83	58
Viable:non viable tumor	0.00	0.00	0.00	11
Gaussian Naïve Bayes				
Non tumour	0.85	0.84	0.85	107
Non Viable tumor	0.84	0.70	0.76	53
Viable tumor	0.79	0.64	0.70	58
Viable:non viable tumor	0.25	0.73	0.37	11
Logistic regression				
Non tumour	0.90	0.91	0.91	107
Non Viable tumor	0.86	0.91	0.88	53
Viable tumor	0.77	0.86	0.81	58
Viable:non viable tumor	0.00	0.00	0.00	11

From the table it can be observed that all classifiers achieved strong performance for majority classes. However combined class “Viable/non viable” is too small for the model to learn pattern, so precision, recall an f-score are 0.00 for almost all algorithms. Logistic regression and SVM performed best overall. Class imbalance significantly affected minority class prediction.
Another table 2 below shows the accuracy of classifiers and the misclassified samples.

Table 2: Accuracy table of classifiers

Classifier	Accuracy	Misclassified
SVM	0.85	34
kNN	0.84	37
Random forest	0.84	36
Gaussian Naïve Bayes	0.75	57
Logistic Regression	0.84	34

Table2 above shows the accuracy of each classifier and misclassified counts of each classifier. SVM has the highest accuracy of 0.85 with 34 misclassified samples, followed closely by Logistic Regression. kNN and random forest showed comparable performance with 0.84 accuracy. Gaussian Naïve Bayes performed the worst with only 0.75 accuracy and 57 misclassifications, indicating that the dataset does not satisfy the independence assumptions required by Naïve Bayes. kNN and Random Forest showed comparable performance with 0.84 accuracy.

Conclusion

In today's era, digital board has plenty of data such as internet of things mobile data social media data health data etc. data is in every form that textual data, tabular data, image, videos etc. For analysing the data intelligently and to predict useful information, the knowledge of artificial intelligence, particular machine learning is essential. The paper used 5 classification algorithms to predict the cancer cases form the dataset of MRI scans of osteosarcoma. The performance comparision of 5 classification algorithm shows significant variation in the ability to classifiy tumor categories. The classifiers demonstated consistently high performance for the majority classes, particularly when the number of samples in the class were high. SVM and Logistic regression achieved the highest F1 score across these major classes. However, for the minor class, all classifiers failed to predict accurately. this highlights the impact of class imbalance, where insufficient sample size prevents the model from learning reliable patterns.

Overall it was concluded that linear models (SVM and Logistic Regression) were better suited for osteosarcoma prediction than distance based or probabilistic models. The imbalance of classes remains the primary limitation in the study. Future work should focus on balancing the dataset and exploring advanced deep learning approaches to improve minority class recognition.

REFERENCES

[1] Sarker IH. “Machine Learning: Algorithms, Real-World Applications and Research Directions”, SN Computer Science. 2021;2(3):160. doi: 10.1007/s42979-021-00592

[2] A. Verma, C. K. Shah, V. Kaur, S. Shah and P. Kumar, "Cancer Detection and Analysis Using Machine Learning," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, pp. 1-5, doi: 10.1109/ICCSEA54677.2022.9936457.

[3] Shaikh FJ, Rao DS. “Prediction of cancer disease using machine learning approach”, Mater Today Proc 2022;50:40–7

[4] Kumar Yogesh, Gupta Surbhi, Singla Ruchi, Hu Yu-Chen. “A systematic review of artificial intelligence techniques in cancer prediction and diagnosis”, Arch Computer Methods Eng 2022;29:2043–70.

[5] Ahmed, M.A., Abdel Moety, A. & Soliman, A.M.A. “Predicting cancer risk using machine learning on lifestyle and genetic data”. Sci Rep 15, 30458 (2025). https://doi.org/10.1038/s41598-025-15656-8

[6] Mohammad Hossein Sadeghi, Sedigheh Sina, Hamid Omidi, Amir Hossein Farshchitabrizi, Mehrosadat Alavi, “Deep learning in ovarian cancer diagnosis: a comprehensive review of various imaging modalities” Pol J Radiol 2024; 89: e30-e48, DOI: https://doi.org/10.5114/pjr.2024.134817

[7] F.J. Shaikh and D.S. Rao, “Predication of cancer disease using machine learning approach”, Materials Today: Proceedings, Volume 50, Part 1, 2022, Pages 40-47https://doi.org/10.1016/j.matpr.2021.03.625

[8] Seeta Devi1, Ruchika Kaul Ghanekar, Jayshree Ashok Pande, Dipali Dumbre, Ranjana Chavan, Harshita Gupta, “Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models” Asian Pacific Journal of Cancer Prevention, Vol 25

[9] Kamta Nath Mishra , Alok Mishra , Soumya Ray, Anjali Kumari, Saad Misbah Waris, “Enhancing cancer detection and prevention mechanisms using advanced machine learning approaches” Informatics in Medicine Unlocked 50 (2024) 101579

[10] Kourou K, Exarchos KP, Papaloukas C, Sakaloglou P, Exarchos T, Fotiadis DI. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Comput Struct Biotechnol J. 2021 Oct 6;19:5546-5555. doi: 10.1016/j.csbj.2021.10.006. PMID: 34712399; PMCID: PMC8523813.

[11] Pereira HM, Leite Duarte ME, Ribeiro Damasceno I, de Oliveira Moura Santos LA, Nogueira-Barbosa MH. Machine learning-based CT radiomics features for the prediction of pulmonary metastasis in osteosarcoma. Br J Radiol. 2021 Aug 1;94(1124):20201391. doi: 10.1259/bjr.20201391. Epub 2021 Jun 19. PMID: 34111978; PMCID:

[12] Varlamova, E. V., Butakova, M. A., Semyonova, V. V., Soldatov, S. A., Poltavskiy, A. V., Kit, O. I., & Soldatov, A. V. (2024). Machine Learning Meets Cancer. Cancers, 16(6), 1100. https://doi.org/10.3390/cancers16061100