Grenze International Journal of Engineering and Technology, June Issue

Traditional SVM, KNN and Random Forest models Comparison on Skin Cancer Dataset

Kirti Wanjale[1], Kaustubh Chaudhari[2], Atharva Choudhari[3], Vikas Maral[4], Achal Ghadge[5] and Sanyukta Deshmukh[6]
[1-6]Computer Engineering, Vishwkarma Institute of Technology, Pune, India,
Email: kirti.wanjale@vit.edu

***Abstract*— A comparative study was carried out to assess the performance of three machine learning models, including Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Random Forest (RF), in classifying skin cancers using a standard dataset. Classification performance is evaluated by various metrics: accuracy, precision, recall, and others, including analysis of computational efficiency based on training time, resource consumed in training, and calculation time. A holistic assessment of these proposes presents merits and demerits of each of these models. Thus, it helps in selecting the most useful algorithm for skin cancer detection. These deliverables aim to increase the comprehension of machine learning applications in distinguishing between malignant melanomas and benign cases.**

***Index Terms*— Random Forest, KNN, Support Vector machine, Supervised learning, Decision Tree.**

I. INTRODUCTION

The research work aimed to explore a detailed analysis of machine learning based skin cancer classification. The study is mainly focused on comparison of various supervised learning models particularly Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbours (KNN). It defined its performance assessment in terms of training 80% on a dataset and then testing on the remaining 20%, ensuring that all possible cases are embedded in the investigation to enable accurate diagnosis of skin cancer using these systems.

Here, we have used various types of machine learning algorithms for solving problems related to the detection of different types of skin cancers, such as melanoma, basal cell carcinoma, and squamous cell carcinoma. These algorithms include 'Traditional Random Forest,' 'Balanced Random Forest,' 'Extra Trees,' 'AdaBoost,' 'Bagging RF,' 'Standard KNN,' 'Weighted KNN,' 'Distance-Weighted KNN,' 'Ball Tree KNN,' 'Minkowski Distance KNN,' 'KD-Tree KNN,' 'Brute Force KNN,' 'Linear SVM,' 'Non-Linear SVM (RBF Kernel),' 'Non-Linear SVM (Polynomial Kernel),' 'Multi-Class SVM,' 'Probabilistic SVM,' 'Weighted SVM,' and 'SMO SVM.' We mainly focused on comparing these models based on their accuracy measures, but other calculated values, such as precision and recall, have also been considered, with computational efficiency being one of them. This research will look at different models and how well they perform based on things like accuracy, precision, recall etc. In addition, we will examine each algorithm's efficiency in terms of training time or resource usage which could help us understand where these methods can be used in real world situations best.

Grenze ID: 01.GIJET.11.2.247_20

So by doing this kind of comparison there might come some new knowledge about what works better than others when it comes to machine learning applied on skin cancer detection; therefore also improving diagnostic accuracy within healthcare systems worldwide.

II. Related Work

A Weighted KNN Approach to Estimate Missing Values: In this paper, they introduce a new method of treating missing values using a weighted KNN strategy. The similarity among distance metric dimensions is evaluated using Grey Relational Grade and SVM Weight-based weights. It is examined on the Fisher Iris data set, showing better performance by blending kernels of polynomials and radial bases than using a mixture kernel. However, it is only useful when dealing with continuous data types while considering how to tackle missing data entries in mixed type environments where there are both numerical and categorical rows in addition to future work/frameworks for areas containing missing values in numerical datasets or categorical data [2].
A Precise Breast Cancer Detection Approach Using Ensemble of Random Forest with AdaBoost: In the paper which discusses employing machine learning classifiers for precise medical decision-making, in particular regarding binary classification of breast tumor (malignant or benign), we note that although there are high incidences of breast cancer it can be cured once detected early enough. In terms of classification the Random Forest and AdaBoost structures revealed encouraging outcomes, implying that prognosis using machine learning outmatches traditional approaches. Consequently, this technique is projected to greatly facilitate in the detection of breast cancer [4].
An Advanced k Nearest Neighbour Classification Algorithm Based on KD-tree: In order to reduce its excessively high computation cost the KNN algorithm described in the paper was enhanced with a KD-tree. Experiments with various standard datasets illustrated the improvement in the search time efficiency of the algorithm. One major disadvantage, however, of the KD-tree is its great space complexity. In future works, it will be planned to add attribute importance weights to the KNN-KD-tree to increase accuracy. Further, different distance metrics for accuracy enhancement will be developed, and further research on diverse distance formulas on a wider range of datasets will be carried out [3].
Marriage Recommendation Algorithm Based on KD-KNN-LR Model: The research explores machine learning's role when it comes to matching up unwed and divorced people. It introduces a new algorithm which is KD-KNN-LR. Due to the growing number of single and divorced people in China, it is important that there are efficient ways to match them up in marriage. KD-KNN-LR method unites LR and KD-KNN approaches to make personalized recommendations according to user characteristics and preferences [1].
Identification of disturbance sources based on random forest model: In this paper they introduce a way to detect disturbances in power quality using random forests. It chooses which indices will be analysed; extracts temporal and statistics properties; balances data sets; trains trees; tunes parameters based on OOB estimates; and finally reaches an optimal level for discriminating between railway related disturbances (that occur due to trains passing through), wind farms generation (that generate electricity) or solar panels installations (sun panels) [6].

III. Dataset

Here in this research, data from the HAM10000 dataset has been used, which is a very important dataset on obtaining metadata that is related to dermatological images depicting skin lesions. This dataset has a lot of important features for classifying and predicting skin cancer. Each record in the dataset can be described with several attributes: e.g., a unique lesion ID, which marks individual skin lesion entries associated with its corresponding image id; diagnosis (dx) with regard to whether and lesion is benign or malignant and as the target variable for classification; diagnosis type (dx type-mostly in terms of histopathology, reference standard). Other information is also included in the dataset, like the age at diagnosis, which can have a part in determining the risk of skin cancer. Sex means the gender of the patient and localization expresses the anatomical location of the lesion at the patient's body-partitioned hand or ear, for example. This is the metadata that will support the creation of predictive models which would estimate the malignancy of skin lesions according to these characteristics and that would be the foundation for the skin lesion classification task of this research.

IV. Methodology

A. SVM models

Adapted and trained various support vector machine (SVM) models for the detection of skin cancer. The SVM variations which were configured included linear SVM, non-linear SVM with RBF kernel, non-linear SVM with

polynomial kernel, multi-class SVM, probabilistic SVM, weighted SVM, and SMO SVM. Each model was tweaked in certain specific ways. The adjustments focused on kernel selection, handling of imbalanced classes, and activating probability estimates. A standard training regimen was carried out. Here there were also some steps of data pre-processing, that is, among other things, scaling and encoding of categorical variables. The performance of every model had to be measured post-training.
This evaluation involved, for instance, making forecasts about new observations while tracking key figures like accuracy, precision, recall, and F1 score. Furthermore, scatter plots were used together with other graphical presentations so as to compare expected outcomes against real ones. This way we were able to get a full picture on how well each SVM model could classify cases of skin cancer on the basis of results got from visual aids.

B. Linear SVM

Utilized a basic linear Support Vector Machine classifier (SVC(kernel='linear')) with default configurations to achieve skin cancer classification.
Performed data pre-processing like dealing with missing observations, scaling features with respect to variables and encoding categories.
Learned linear boundary lines between different classes of skin cancer by training the Linear SVM model (svm_linear) on a processed training set (X_train and y_train).

C. Non-Linear SVM with RBF Kernel

Used a Radial Basis Function (RBF) kernel (SVC(kernel='rbf')) in an SVM model to handle non-linear patterns in skin cancer classification.
Increase input data complexity to enable SVM acquisition of complex relationships through feature engineering techniques which have been applied.
Learn non-linear decision boundaries through the Non-Linear SVM (RBF Kernel) model (svm_rbf) training using the transformed training data.

D. Probabilistic SVM

Arranged an SVM model that will produce probable outputs (SVC (probability = True)) that are necessary for estimating the class probabilities in tasks related to skin cancer classification.
Here we implemented calibration either through probability scaling or isotonic transformation in order to map the SVM's outputs into reliable probability estimates hence improving classification accuracies.
To enable trustworthy probabilistic predictions of skin cancer classes, the likelihood ratios of the probabilistic SVM model were computed afterwards it was trained using pre-processed training data that had been provided during the process.

E. SMO SVM

To handle linearly separable cases in skin cancer classification, trained an SVM model using a linear kernel (SVC(kernel='linear', C=1.0)) through Sequential Minimal Optimization (SMO) algorithm.
Performed feature selection or dimensionality reduction to help the SVM deal with high-dimensional data more effectively. By optimizing the linear decision boundaries for accurate classification of skin cancer instances, trained the SMO SVM model (smo_svm) with pre-processed training data.

F. Random forest models

Traditional Random Forest:
Random Forest generally consists of hundreds to thousands of decision trees, usually trained using substantial data. A subset of some particular training data is learnt by each decision tree and contributes to ensemble's predictive power.

G. Balanced Random Forest

Our focus was to address class imbalance while still training similarly to the traditional model. This model employs an equivalent number of decision trees that the original Random Forest model uses thus ensuring the ensemble's robustness.

H. Random Forest with Extra Trees

Compared to conventional Random Forests, randomization was increased in the growth of trees to train this model.
This model also utilizes a substantial number of decision trees, often ranging from hundreds to thousands

I. AdaBoost (Adaptive Boosting)

It is process of training model with many weak learners which in AdaBoost are decision trees.
Thus, an exact multiple of decision trees such as its boosting iterations or even data complexity can also be included in AdaBoost.

J. Bagging Random Forest

Implemented Bagging to train multiple Random Forest models work together, with each model consisting of many decision trees.
The performance of the model as a whole is due to the combined predictions of these decision trees.

K. KNN models

Investigated the K-Nearest Neighbours (KNN) algorithm using different models that exhibit distinct attributes, which are Standard KNN, Weighted KNN, Distance-Weighted KNN, Ball Tree KNN, Minkowski Distance KNN, KD-Tree KNN, and Brute Force KNN. These models serve as good examples of how the algorithm can be adjusted to various data sets and problem domains providing range of alternatives for efficient nearest neighbour search and classification tasks.

L. Weighted KNN

We used KNeighboursClassifier(n_neighbours=5, weights='distance') bearing in mind that near neighbours have a stronger impact on the outcome than distant ones because of weight based on distance.
In this case “reversed distance-based” distance means that the emphasis placed by each neighbour while predicting will change according to their closeness or farness to a test point. This causes us to have better resolution of the classes by giving priority to nearer compared with remote data points in making decisions in Weighted K-Nearest Neighbour (WKNN).

M. Distance-Weighted KNN

We used KNeighboursClassifier(n_neighbours=5, weights=lambda x: 1 / (x + 1e-10)) with our own distance-weighting function, in which emphasis is put on the significance of nearby neighbours and de-emphasizing those farther away.
The method is like weighted KNN but involves modification to the weighting function that decreases the relevance of far neighbours thereby increasing dependency on adjacent ones (points)"

V. RESULTS

A. SVM models

Best model: Multi-Class SVM

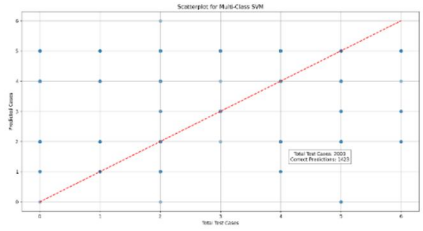

Figure 1: Predictions Multi-class SVM

Least Perfect model: Weighted SVM

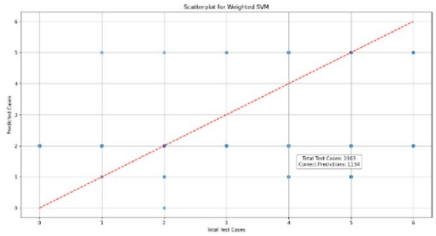

Figure 2: Predictions Weighted SVM

TABLE I: OVERALL ANALYSIS OF ALL MODELS OF RANDOM FORCE

Method	**Accuracy**	**Precision**	**Recall**	**F1score**
Linear SVM	0.7000	0.5457	0.7000	0.6091
Non-Linear SVM (RBF Kernel)	0.7104	0.6254	0.7104	0.6522
Non-Linear SVM (Polynomial Kernel)	0.7089	0.6277	0.7089	0.6501
Multi-Class SVM	0.7104	0.6254	0.7104	0.6522

A. KNN models

Best model: Weighted KNN

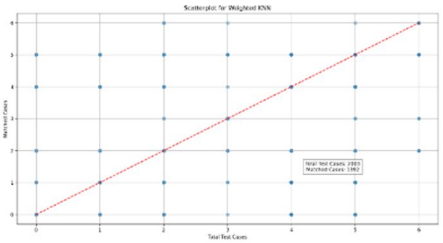

Figure 3: Predictions Weighted KNN

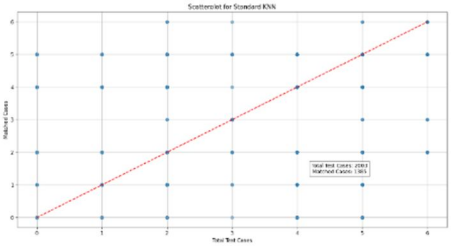

Figure 4: Predictions Standard KNN

Least Perfect model: Standard KNN

TABLE II: OVERALL ANALYSIS OF ALL MODELS OF KNN

Method	Accuracy	Precision	Recall	F1score
Standard KNN	0.691463	0.704594	0.704594	0.704594
Weighted KNN	0.694958	0.704692	0.704692	0.704692
Distance-Weighted KNN	0.694958	0.704692	0.704692	0.704692

B. Random Forest models:

Best model: Bagging Random Forest

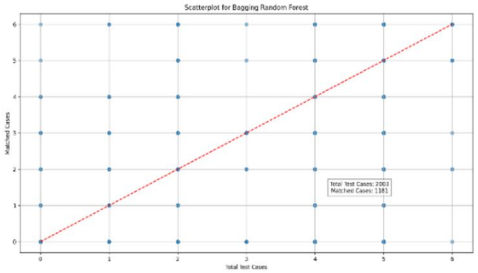

Figure 5: Predictions Bagging Random Forest

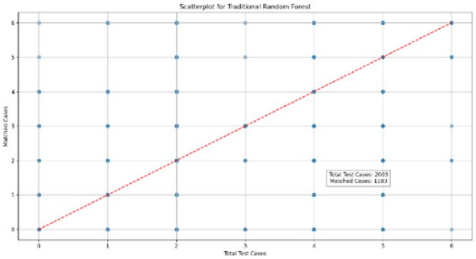

Figure 6: Predictions Traditional Random Forest

TABLE III: OVERALL ANALYSIS OF ALL MODELS OF RANDOM FORCE

Method	**Accuracy**	**Precision**	**Recall**	**F1score**
Traditional Random Forest	0.7498	0.7519	0.5916	0.6450
Balanced Random Forest	0.7498	0.7527	0.5926	0.6459
Extra Trees	0.7434	0.7511	0.5901	0.6438

Comparison of different SVM methods revealed that the Multi-class SVM worked better than other SVM models with its accuracy being 71.04%, an indication of its ability to generalize well on unseen data points while also ensuring that there are minimal errors made during the process. However, Weighted SVM had 56.62% accuracy which was regarded as the lowest among all the SVM models. It clearly shows how much varied SVM types perform when it comes to this kind of work, with the Multi-class SVM emerging as the top performer in this evaluation.

A comparison of different KNN models showed that Weighted K-Nearest Neighbour (KNN) and Distance-Weighted KNNs worked better; they correctly classified 69.49% examples, based on the their nearest neighbours and correct applied weights. In contrast to the rest versions, the Standard model misclassified most of them standing at 69.16% accuracy rate.

VI. Conclusion

Our study showed that ensemble methods like Bagging Random Forest often produce higher accuracy compared to individual classifiers, thereby revealing their potential in tackling the challenges of over-fitting and variance. Therefore, Bagging Random Forest model had an accuracy level of 75.36% which portrays how it can handle different types of data while the SVM model showed less accurate results with a score of 56.62%. Moreover, it was shown that ensemble methods provide superior machine learning models than individual algorithms. Such results reveal their potential in controlling over-fitting and variance when they are used in complex systems. Consequently, Bagging Random Forest model demonstrated an accuracy level of 75.36% which indicates its ability to operate across various data structures whereas the Weighted SVM model exhibited an accuracy of 56.62%.

Going forward, it is very important to keep exploring different machine learning algorithms and techniques all the time. This means tweaking the hyper parameters properly, finding ways to create features as well as using ensemble methods so that you have an improved performance. When choosing models or optimizing them, understanding details such as the distribution of classes in data sets or which are most significant features becomes helpful.

In conclusion, it emphasizes the iterative nature of ML developing and the criticalness of continuous learning for optimal results in predictive modeling tasks.

References

[1] Cai, Z., & Zhang, X. (2020, June). Marriage recommendation algorithm based on KD-KNN-LR model. In 2020 IEEE International Conference on Artificial Intelligence and Computer Application (ICAICA) (pp. 569-572). IEEE.

[2] Sen, S., Das, M. N., & Chatterjee, R. (2016, February). A weighted kNN approach to estimate missing values. In 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 210-214). IEEE.

[3] Hou, W., Li, D., Xu, C., Zhang, H., & Li, T. (2018, December). An advanced k nearest neighbor classification algorithm based on KD-tree. In 2018 IEEE International Conference of Safety Produce Informatization (IICSPI) (pp. 902-905). IEEE.

[4] Rohan, T. I., Siddik, A. B., Islam, M., & Yusuf, M. S. U. (2019, July). A precise breast cancer detection approach using ensemble of random forest with AdaBoost. In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

[5] Sahu, S. K., Pujari, A. K., Kagita, V. R., Kumar, V., & Padmanabhan, V. (2015, December). Gp-svm: Tree structured multiclass svm with greedy partitioning. In 2015 International Conference on Information Technology (ICIT) (pp. 142-147). IEEE.

[6] Hao, Z., Shaohong, L., & Jinping, S. (2011, October). Unit Model of Binary SVM with DS Output and its Application in Multi-class SVM. In 2011 fourth international symposium on computational intelligence and design (Vol. 1, pp. 101-104). IEEE.

[7] Govada, A., Gauri, B., & Sahay, S. K. (2015, August). Centroid based Binary Tree Structured SVM for multi classification. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 258-262). IEEE.

[8] Geng, R., Zhang, D., & Chai, J. (2012, August). Research on grain information classification based on SVM decision tree. In 2012 IEEE International Conference on Granular Computing (pp. 138-141).IEEE.

[9] Arshpreet Kaur, Abhijit Chitre, Kirti Wanjale, Pankaj Kumar, Shahajan Miah, Arnold C. Alguno, "Recognition of Protein Network for Bioinformatics Knowledge Analysis Using Support Vector Machine", BioMed Research International, vol. 2022, Article ID 2273648, 11 pages, 2022. https://doi.org/10.1155/2022/2273648

[10] Yuan, D., Huang, J., Yang, X., & Cui, J. (2020, November). Improved random forest classification approach based on hybrid clustering selection. In 2020 Chinese Automation Congress (CAC) (pp. 1559-1563). IEEE.

[11] Zhan, Y., Dai, S., Mao, Q., Liu, L., & Sheng, W. (2015). A video semantic analysis method based on kernel discriminative sparse representation and weighted KNN. The Computer Journal, 58(6), 1360-1372.