IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 10 Issue 01, January 2023
ISSN (Online) 2348 – 7968 | Impact Factor – 6.72
www.ijiset.com

Comparative Analysis of Machine Learning Models in the Histopathologic Grading of Oral Squamous Cell Carcinoma

Shwethal Sayeeram Trikannad[1], Noel Pereira[2]

[1] SDM College of Dental Sciences and Hospital, Sattur, Dharwad, Karnataka, India

[2] Manipal Institute of Technology, Manipal, Karnataka, India

Abstract

One of the most common malignancies in the head and neck region is the Oral Squamous Cell Carcinoma (OSCC). Its high mortality and morbidity coupled with a low survival rate make this cancer a pressing concern the world over [1]. TNM staging standards (8th edition, 2017) by the American Joint Committee on Cancer (AJCC) and the International Union against Cancer (UICC) and World Health Organization (WHO) histopathologic grading (4th edition, 2017) are a key step in management and treatment of OSCC. Machine learning has applications in varied domains of OSCC [2]. We present machine learning approaches including Random Forest, Logistic Regression, Adaboost Classifier, Gradient Boosting and Extra Trees Classifier to analyze, correlate and subsequently, predict the histopathological grading of OSCC utilizing multiple data obtained from The Cancer Genome Atlas (TCGA) database [3]. The results, post data wrangling were found to be 62.5%, 59.375%, 59.375%, 53.125% and 54.6875% respectively.

Keywords: *Oral Squamous Cell Carcinoma, TNM staging, Histopathologic grading, Machine learning.*

1. Introduction

Cancer is one of the top causes of death globally. Oral cancer especially, is the leading cause of mortality among all other oral diseases. The yearly global incidence as of 2020 according to the Global Cancer Observatory (GCO) was 377,713 cases with the majority from Asia with 248,360 followed by Europe with 65,279 and lastly, North America recorded 27,469 cases [4]. The risk factors range from tobacco and alcohol to carcinogenic viruses like the human papillomavirus (HPV), oro-dental factors and dietary deficits [5]. Existence of any pre malignant lesions and conditions like oral lichen planus (OLP), leukoplakia, erythroplakia and oral submucous fibrosis (OSMF) substantiate the risk of developing oral cancer [6].

Adenocarcinoma (ADC) and Squamous cell carcinoma (SCC) constitute the two major types of carcinomas affecting epithelial tissue. SCC can be found in a number of sites such as the head and neck region, lung, prostate, skin, cervix and vagina [7]. Early detection and timely management of any malignancy helps in decreasing morbidity and mortality and increasing long-term survival rates. The next step after detecting a malignancy is staging and grading it with the help of two important standardized classification systems namely, the TNM (T-primary tumor size N- lymph node involvement M-distant metastasis) staging and histopathologic grading. These help with appropriate treatment planning, estimating recurrence risk, assessing prognosis and ultimately indicating how successfully can the given patient be cured.

Treatment of OSSC is multimodal involving surgical resection with any required adjuvant therapy like radiotherapy and chemotherapy. The need for adjuvant therapy is decided on the basis of severity of the lesion decided by the TNM staging and histopathological grading. Characteristics taken into consideration are the histopathology report depicting differentiation, depth of invasion, metastasis, nodal status among others; location of primary tumor and presence of vital structures nearby [8].

With machine learning emerging as an essential aid in the areas of cancer prognosis, diagnosis and treatment planning, we hereby introduce this research paper which compares the accuracy in predicting the histopathological grading when TNM

staging along with certain other characters are loaded into a variety of machine learning models including Random Forest, Logistic Regression, Adaboost Classifier, Gradient Boosting and Extra Trees Classifier, using data acquired from TCGA database.

2. Literature Review

Random Forest based algorithms used to predict high risk identify oral premalignant lesions had a positive predictive value of 74% and negative predictive value of 76% [9]. The first study that implemented Artificial Neural Networks (ANN) with a web-based prognostic tool to predict the recurrences in oral tongue squamous cell carcinoma yielded an accuracy of 92.7% for ANN and 88.2% for the web-based prognostic tool [10]. In another study the Decision tree classifier had the highest accuracy to predict the five-year recurrence free survival rate for oral squamous cell cancer using clinical and histopathologic data with 76% followed by Logistic Regression with 60% [11]. A predictive model built with XGBoost architecture to predict occult nodal metastasis in early oral squamous cell carcinoma using multi-institutional clinicopathological data had a sensitivity of 91.7%, specificity of 72.6%, positive predictive value of 39.3% and negative predictive value of 97.8% [12].

Most studies utilized machine learning algorithms to predict incidence of oral cancer, potential of development of oral cancer in premalignant lesions, risk of recurrence, risk of metastasis, survival rates of patients. This paper however, utilizes clinical data and TNM staging to predict histopathological grading thus being an essential aid to clinicians in formulating a case appropriate treatment protocol rapidly, especially in the cases of aggressive tumors where time is of the essence.

3. Theoretical Review

3.1 Oral squamous cell carcinoma

Oral squamous cell carcinoma is derived from the oral mucosal epithelium and is usually found on the lateral borders of the tongue, lips, buccal mucosa, hard palate and retromolar trigone area. Carcinoma associated with carcinogens like alcohol and tobacco is seen in Southeast Asia and Australia [13]. Whereas, Human Papilloma Virus (HPV) related infections are the major cause of this malignancy along with other squamous cell carcinomas in the head and neck region in USA and Western Europe [14][15]. Males are at a higher risk of developing OSCC than females. The risk factors for OSCC are tobacco consumption including smokeless tobacco, areca nut, betel quid and alcohol consumption, environmental pollutants and viruses like HPV and Epstein-Barr Virus (EBV) and genetic factors [16].

The progression of healthy epithelium to one that is malignant undergoes a number of histopathological stages namely, cell hyperplasia (increase in number of cells), cell dysplasia (increase in abnormal characters of cells), carcinoma in situ and frank invasive malignancy [16]. Apart from the etiological agents a number of biomarkers have been identified and associated with cancer stem cells that may originate from mature progenitor pluripotent stem cells like CD44(type 1 transmembrane glycoprotein), CD133(prominin 1) and ALDH1(aldehyde dehydrogenase 1) [17][18].

Clinically, OSSC presents as a non-healing ulcer and is usually caught in the initial stages as patient is aware of the ulceration. Definitive diagnosis however, is with the help of histopathological study of a biopsy procured from the primary lesion and neck masses, if any [19]. Histopathologic grading depends on the type of hyperplastic or dysplastic features seen in the biopsied tissue and is the gold standard in the confirmation of presence of malignancy.

Staging according to TNM system is done with the help of extensive head and neck clinical examination along with CT and MRI scans to identify the extent of the lesion. Chest CT scans are conducted to rule out any distant metastases [16].
Management of OSCC depends on the size and grade of tumor, anatomical location, presence of nearby vital structures, nodal involvement and metastases. Conventional treatment involves a combination of radical surgery, radiotherapy and chemotherapy. Occult nodal involvement show better prognostic results with elective neck dissection [20].

3.2 TNM staging

The malignancy staging system called TNM staging assesses the expanse of tumor growth in the whole body with T standing for primary tumor size, N representing the involvement of regional lymph nodes and M is the distant metastases. This system considers only the anatomic involvement and does not take into consideration any other prognostic factors [21].

The TNM staging was first published in the year 1968 by the International Union Against Cancer (UICC) which was followed by the American Joint Committee on Cancer (AJCC) in the year 1977 by publishing its first staging manual. The 8th edition of the UICC and AJCC staging manual was released in 2017 which included two major changes, one it introduced tumor depth of invasion (DOI) in the T stage and extracapsular spread in the N stage [22][23].

3.3 Histopathologic grading

Broders [24] introduced the histopathological grading based on SCC of the lip. More grading systems were suggested by Jakobsson et al. [25], Anneroth et al. [26] and Bryne et al. [27]. The World Health Organization (WHO) is based on the Broders criteria and has three grades of OSCC namely well-, moderately-, and poorly-differentiated [28].

Specimen for the grading is taken as a wedge-shaped tissue from the suspected lesion including the healthy tissue for comparison. Grading is carried out with the help of the number of undifferentiated cells seen with hyperplastic and/or dysplastic features. The more the cells the higher the grade and invariably, the more aggressive the cancer.

3.4 Random Forest

Random forest, practically, is a model built atop the bedrock idea of decision trees, which are supervised learning models that can be utilized for both classification and regression tasks. Decision trees function on the selection of an individual feature, labelling it as the most important and then generating outputs on the basis of it [29]. Random forest develops on the idea of decision trees by randomizing feature selection and then generating multiple trees to produce the outputs, which are then bundled to give a final output. It gets it coinage from feature randomizing and usage of multiple trees [30].

3.5 Logistic Regression

Despite the irony of the name, logistic regression finds its scope of utilization, in tasks involving classification. It stands as the most efficient model for binary or linear classification method. It can be generalized over a large number of features and produces high accuracy models when the features tend to be independent of each other. When there are more than two discrete outcomes, it is called multinomial logistic regression.

3.6 Adaboost Classifier

Adaboost classifier, exists as an amalgamation of several weaker models. It first exists as a boosting algorithm which entails the collection, optimization and combination of several less accurate algorithms, to generate a stronger one. The most commonly utilized algorithms are decision trees with a single split, named stumps. To it, Adaboost adds higher weightage on existing features that are difficult to interpret, and tends to ignore the easily understandable features. It finds use in both classification and regression tasks [29].

3.7 Gradient Boosting Classifier

Gradient Boosting Classifier, works on the basis of gradient descent and boosting. In a myriad of ways, it is similar to Adaboost, in the utilization of several weaker models, to generate a strong learner. While Adaboost use stumps, the weak

learner trees in Gradient Boosting, are constructed with greedy algorithms on the basis of split points and purity scores, and since they utilize a generic algorithm, they are more flexible than Adaboost Classifiers [31].

3.8 Extra Trees Classifier

Extra Trees is similar to Random Forest, in that it generates a variety of trees and engages in the splitting of nodes using a randomized set of features. It is different from random forest due to the absence of observation bootstrapping, that is, it samples the data without replacement and the tree nodes are split randomly as opposed to Random Forests which find the best splits [30].

4. Research Methodology

The dataset was contributed by Liu Fangzhou (2020) and was utilized in a research paper discussing a novel 7 immune related gene prognostic model for oral cancer using the TCGA database [3]. The data consisted of clinicopathologic and genetic information of 213 anonymous patients with TCGA identification numbers, of which only the former was used. The clinical characters were age, gender, T score, N score, M score, TNM stage and histopathologic grade.

T scores comprised of the least unknowns with others being either Tx- primary tumor cannot be assessed, T1- tumor size < 2cm diameter and depth of invasion (DOI) < 5mm, T2- tumor size < 2cm diameter and DOI > 5mm but < 10mm or tumor size 2-4 cm in diameter and DOI< 10 mm, T3 – tumor size > 4cm diameter or any tumor with DOI> 10mm or peripheral neural invasion of large caliber nerves (greater than or equal to 0.1mm), T4a- moderately advanced local disease or T4b- Very advanced local disease; tumor invades masticator space, pterygoid plates or skull base and/or encases the internal carotid artery.

M scores had the most unknowns with some Mx- metastasis cannot be measured and M0- absence of distant metastasis. N scores had a few unknowns with others being Nx- regional lymph nodes cannot be assessed, N0- No regional lymph node metastasis, N1- single ipsilateral lymph node metastasis less than or equal to 3cm in diameter with no extranodal extension (ENE), N2a- single ipsilateral lymph node metastasis >3cm but < 6cm with no ENE, N2b- multiple ipsilateral lymph node metastases none > 6cm with no ENE, N2c- bilateral or contralateral lymph node metastases with none > 6cm with no ENE, N3 – single lymph node metastasis >6cm with no ENE or metastasis in single ipsilateral lymph node with ENE present or multiple ipsilateral , contralateral or bilateral nodes of any size with ENE present.

TNM stages comprised of a few unknowns with others labelled as stage I - T1 N0 M0, stage II - T2 N0 M0, stage III - T2 Ni M0 or T2 N1 M0 or T3 N0,N1 M0, stage IVa - T4 N0,N1 M0 or stage IVb - any T, N2,N3 M0 or any T, any N, M1. Histopathologic grading ranged from Gx- grade cannot be assessed, G1- well differentiated (<25% undifferentiated cells), G2- moderately differentiated (<50% undifferentiated cells) and G3- poorly differentiated (<75% undifferentiated cells).

After the acquisition of the dataset, the imported data went through the data wrangling pipeline which included identification and elimination of missing values and NaN values, data type conversions and label encoding, which is essentially the conversion of categorical data points into its numerical counterparts making training easier. This resulted in a data frame in which each feature was numerical. The dataset was then split into 70% training and 30% testing set which were then fit and trained into the individual models, and then tested on the testing set.

5. Results

The results, post data wrangling and passing the cleaned data to the models is displayed in Table 1.

ISSN (Online) 2348 – 7968 | Impact Factor – 6.72
www.ijiset.com

Table 1: Accuracy of each machine learning model

Serial Number	***Model***	Accuracy
1	Random Forest	62.5%
2	Logistic Regression	59.375%
3	Adaboost Classifier	59.375%
4	Gradient Boosting Classifier	53.125%
5	Extra Tree Classifier	54.6875%

The heat maps predicting the false positives, false negatives, true positives and true negatives of all the models are depicted below in the following order Figure 1., Figure 2., Figure 3., Figure 4. and Figure 5 corresponding to Random Forest, Logistic Regression, Adaboost Classifier, Gradient Boosting Classifier and Extra Tree Classifier respectively.

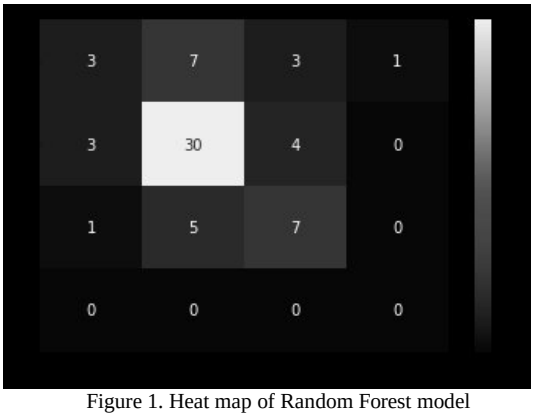

Figure 1. Heat map of Random Forest model

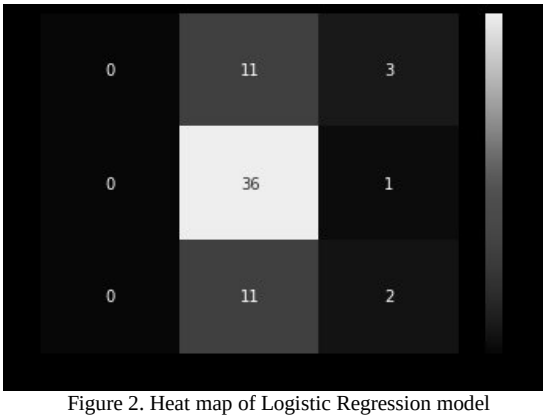

Figure 2. Heat map of Logistic Regression model

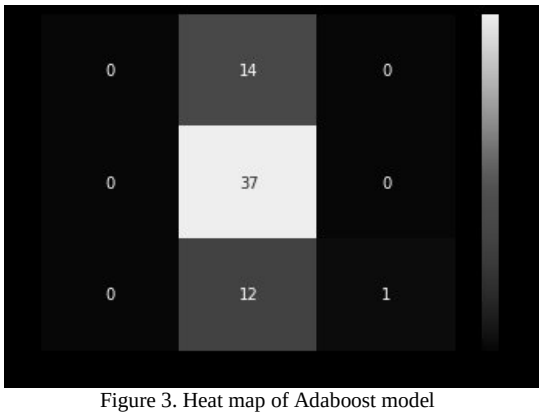

Figure 3. Heat map of Adaboost model

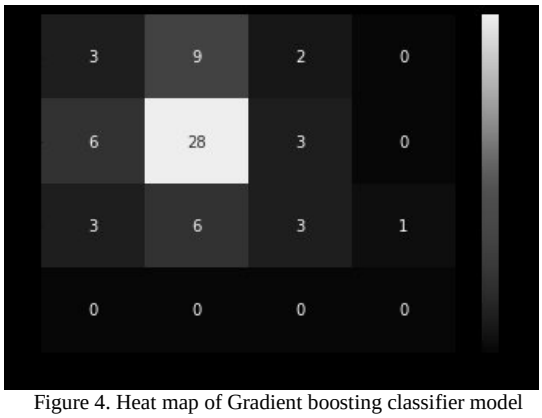

Figure 4. Heat map of Gradient boosting classifier model

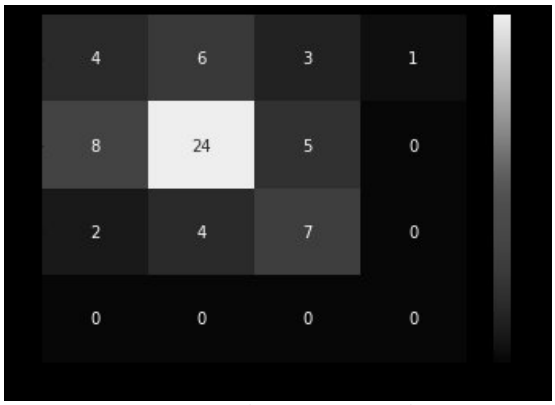

Figure 5. Heat map of Extra Tree Classifier model

6. Discussion and Conclusion

As evidenced, machine learning models can be a very useful tool to help clinicians battle cancer. However, there is room for improvement. Some ways would be to include the trials of better machine learning models like XGBoost or better optimization techniques and tweaking the bias-variance trade-off. Appending the dataset with more usable data and less unknowns and better hyperparameter tuning would also help achieve higher accuracy and ultimately, create more robust algorithms.

Acknowledgements

We would like to thank the Almighty for all the blessings showered on us. We would also like to thank our respective institutions-SDM College of Dental Sciences and Hospital and Manipal Institute of Technology for providing us with countless opportunities and the necessary knowledge to utilize them. Finally, we are very grateful to our family, for being an unerring source of love, support and inspiration.

References

[1] Pulte D, Brenner H. “Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis.” Oncologist 2010; 15: 994-100
[2] López-Cortés XA, Matamala F, Venegas B, Rivera C. “Machine-Learning Applications in Oral Cancer: A Systematic Review.” Applied Sciences. 2022; 12(11):5715
[3] Liu, Fangzhou, “Data for: Development of a novel 7 immune-related genes prognostic model for oral cancer: a study based on TCGA database.”, Mendeley Data, V1, 2020
[4] “Oral cancer - the fight must go on against all odds….” Evid Based Dent 23, 4–5 (2022).
[5] Ram H, Sarkar J, Kumar H, Konwar R, Bhatt ML, Mohammad S. “Oral cancer: risk factors and molecular pathogenesis.” J Maxillofac Oral Surg. Jun;10(2):132-7, 2011 doi: 10.1007/s12663-011-0195-z. Epub 2011 Apr 22. PMID: 22654364; PMCID: PMC3177522.
[6] Yardimci G, Kutlubay Z, Engin B, Tuzun Y. “Precancerous lesions of oral mucosa.” World J Clin Cases. Dec 16;2(12):866-72, 2014 doi: 10.12998/wjcc.v2.i12.866. PMID: 25516862; PMCID: PMC4266835
[7] Ali Hasan Md. Linkon, Md. Mahir Labib, Tarik Hasan, Mozammal Hossain, Marium-E- Jannat,“Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study”, Informatics in Medicine Unlocked, Volume 24,2021,100582,ISSN 2352-9148
[8] Alhadi Almangush, Antti A. Mäkitie, Asterios Triantafyllou, Remco de Bree, Primož Strojan, Alessandra Rinaldo, Juan C. Hernandez-Prera, Carlos Suárez, Luiz P. Kowalski, Alfio Ferlito, Ilmo Leivo, “Staging and grading of oral squamous cell carcinoma: An update”, Oral Oncology, Volume 107,2020,104799, ISSN 1368-8375
[9] Baik, J., Ye, Q., Zhang, L. et al. “Automated classification of oral premalignant lesions using image cytometry and Random Forests-based algorithms.” Cell Oncol. 37, 193–202 (2014).

[10] Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, Mäkitie AA, Salo T, Leivo I, Almangush A. "Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool." Virchows Arch. 2019 Oct;475(4):489-497. doi: 10.1007/s00428-019-02642-5. Epub 2019 Aug 17. PMID: 31422502; PMCID: PMC6828835.
[11] Alkhadar, Huda; Macluskey, Michaelina; White, Sharon; Ellis, Ian; Gardner, Alexander "Comparison of machine learning algorithms for the prediction of five-year survival in oral squamous cell carcinoma." Journal of Oral Pathology & Medicine, 2020, jop.13135–. doi:10.1111/jop.13135
[12] Farrokhian N, Holcomb AJ, Dimon E, Karadaghy O, Ward C, Whiteford E, Tolan C, Hanly EK, Buchakjian MR, Harding B, Dooley L, Shinn J, Wood CB, Rohde SL, Khaja S, Parikh A, Bulbul MG, Penn J, Goodwin S, Bur AM. "Development and Validation of Machine Learning Models for Predicting Occult Nodal Metastasis in Early-Stage Oral Cavity Squamous Cell Carcinoma". JAMA Netw Open., Apr 1;5(4):e227226, 2022 doi: 10.1001/jamanetworkopen.2022.7226. PMID: 35416990; PMCID: PMC9008495.
[13] Blot, W. J. et al. "Smoking and drinking in relation to oral and pharyngeal cancer." Cancer Res. 48, 3282–3287 (1988)
[14] Jiang, H. et al. "Can public health policies on alcohol and tobacco reduce a cancer epidemic? Australia's experience." BMC Med. 17, 213 (2019).
[15] Mehanna, H. et al. "Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer–systematic review and meta-analysis of trends by time and region." Head Neck 35, 747–755 (2013).
[16] Johnson, D.E., Burtness, B., Leemans, C.R. et al. "Head and neck squamous cell carcinoma." Nat Rev Dis Primers 6, 92 (2020).
[17] Faber, A. et al. "CD44 as a stem cell marker in head and neck squamous cell carcinoma." Oncol. Rep. 26, 321–326 (2011).
[18] Yu, S. S. & Cirillo, N. "The molecular markers of cancer stem cells in head and neck tumors." J. Cell Physiol. 235, 65–73 (2020).
[19] Pynnonen, M. A. et al. "Clinical practice guideline: evaluation of the neck mass in adults." Otolaryngol. Head Neck Surg. 157, S1–S30 (2017)
[20] D'Cruz, A. K. et al. "Elective versus therapeutic neck dissection in node-negative oral cancer." N. Engl. J. Med. 373, 521–529 (2015).
[21] S.G. Patel, W.M. Lydiatt "Staging of head and neck cancers: is it time to change the balance between the ideal and the practical?" J Surg Oncol, 97 (2008), pp. 653-657
[22] M.B. Amin, S. Edge, F. Greene, D.R. Byrd, R.K. Brookland, M.K. Washington, et al. AJCC Cancer Staging Manual (8th edition), Springer, New York (2017)
[23] Brieleyt JD, Gospodarowicz MK, Wittekind C. "TNM classification of malignant tumors", 8th Edition. Willey Blackwell; 2017, 272p ISBN 978-1-119-26357-9
[24] A.C. Broders "Squamous cell cancer of the lip: a study of five hundred and thirty-seven cases" JAMA, 74 (1920), pp. 656-664
[25] P.A. Jakobsson, C.M. Eneroth, D. Killander, G. Moberger, B. Martensson "Histologic classification and grading of malignancy in carcinoma of the larynx" Acta Radiol Ther Phys Biol, 12 (1973), pp. 1-8
[26] G. Anneroth, J. Batsakis, M. Luna "Review of the literature and a recommended system of malignancy grading in oral squamous cell carcinomas"Scand J Dent Res, 95 (1987), pp. 229-249
[27] M. Bryne, H.S. Koppang, R. Lilleng, A. Kjaerheim "Malignancy grading of the deep invasive margins of oral squamous cell carcinomas has high prognostic value" J Pathol, 166 (1992), pp. 375-381
[28] El-Naggar AK, Chan J, Grandis J, Takata T, Slootweg P. "WHO classification of head and neck tumours." 4th ed.; 2017. p. 105–111
[29] Noel Pereira, "Novel Technology to Revolutionize the Process of Jury Selection and Voir Dire in Criminal Cases," 3 (4) IJLSI Page 689 - 698 (2021)
[30] Medium, https://medium.com/@namanbhandari/extratreesclassifier-8e7fc0502c7, 2018
[31] Educba, https://www.educba.com/gradient-boosting-vs-adaboost/ ,2022

Shwethal Sayeeram Trikannad: Shwethal (B.D.S-2020) works at a private dental clinic who has been accepted into Springer book series. Current research interests include genomics, proteomics and bioinformatics.

Noel Pereira: Noel (BTech- 2021) works as an NLP and SDE at HaiX. Has published two papers in IEEE and one in IJSLI.