Unlocking the Potential of Patient Metadata for Skin Cancer Detection: An AI Framework

Md Shafiqul Islam[1], Gordon Wishart[2], Joseph Walls[3], Per Hall[4], Alba Garcia[1], John Gan[1], Haider Raza[1]

Abstract—**Early detection of suspicious skin lesions can significantly increase the five-year survival rates of the patients. Advancements in computer vision techniques facilitate the use of artificial intelligence (AI) models along with image data for skin cancer detection. However, there is limited work done on skin cancer detection solely based on patient metadata. The 7-point checklist (7PCL) and Williams methods use a limited number of meta-features to calculate skin lesion risk scores and to find a patient at risk of developing skin cancer, respectively. This study attempts to fill the gap and proposes an AI-based framework for classifying skin lesion metadata into binary classes: *Suspicious* vs *Non-suspicious*. The developed framework has been evaluated using real-world skin lesion metadata sourced from a network of private skin diagnostic clinics across the UK. We have collected and analyzed 54,000 skin lesions metadata, from 25,214 patients undergoing teledermatology assessment after clinical examination and imaging, comprising 25 features including patient age, gender, and lesion location. The metadata has been pre-processed through encoding, followed by feature selection using wrapper, Shapley, and Pearson correlation methods. Finally, five different predictive models were utilized and optimized to classify skin lesion metadata into *Suspicious* vs *Non-suspicious* classes. Our proposed approach achieved $83.53(\pm 0.03)\%$ sensitivity in detecting suspicious lesions using only metadata and outperformed the 7PCL and Williams methods. We believe this AI-based framework is unique in classifying skin lesions based solely on metadata and has significant potential to improve the performance of current AI models that are based on image assessment alone.**

I. Introduction

Malignant melanoma is considered one of the fatal types of skin cancer accounting for 90% of deaths among patients with this disease [1]. Delays in early detection of suspicious skin lesions can decrease five-year survival rates by 20% [2], including the United Kingdom (UK) population. The UK follows a two-week wait pathway system, where suspicious lesions of melanoma or squamous cell carcinoma (SCC) should be seen by a specialist within two weeks. These urgent referrals have increased dramatically in recent years (159,430 patients in 2009/2010 to 506,456 patients in 2019/2020), and have significantly contributed to building up healthcare access pressure for timely assessment [3]. Moreover, for non-urgent referrals, such as suspected basal cell carcinoma (BCC), the current waiting time is 18 weeks and only 80% of the patients were seen within this target time frame during 2019/2020. Furthermore, COVID-19 contributed to an increased backlog of non-urgent cases due to cancellations or to accommodate two-week urgent patients. Skin cancer referrals are anticipated to rise in the years ahead because of the ageing population [4] in the UK. In this scenario, artificial intelligence (AI) can emerge as a solution as it has great potential to provide a second opinion about a skin lesion whether it is suspicious or not, and can help in decision-making in the skin cancer assessment pathway.

Due to significant advancements in computer vision (CV) technology, researchers now use image data for skin cancer detection [5]. In the early 2000s, skin cancer was diagnosed using techniques such as the ABCD rule and 7-point checklist (7PCL) before researchers [6, 7] assessed the reliability of introducing CV techniques in skin cancer diagnosis. Later, the availability of open-source datasets and deep learning algorithms such as convolutional neural networks (CNN) were employed for assessing skin lesion images. Esteva et al. [8] compared the performance of CV models versus 21 board-certified dermatologists and achieved a dermatologist-level performance. Conversely, the study in [9] emphasized the importance of the patient's clinical information such as age, gender, and lesion location, and found an overall 7% increase in balanced accuracy with the inclusion of clinical information in the analysis. Similarly, the work in [10] evaluated all combinations of dermoscopic, macroscopic, and clinical metadata (age, gender, and anatomic location) and observed that combining all three yielded the highest overall AUC of 88.80%. The study in [11] evaluated just the clinical information (age, gender, BMI, ethnicity, hypertension, heart disease, and diabetes status) using the National Health Interview Survey (NHIS) data from 450,000 patients between 1997 and 2015 to classify non-melanoma skin cancers against the "never-cancer" skin diseases. They employed a basic feed-forward neural network and achieved an AUC of 81% with 86.2% sensitivity and 62.7% specificity on the validation set.

The previous studies [9, 10, 11] all included a limited set of meta-features (age, gender, anatomic location). Moreover, those studies used metadata along with image data, and there is no mention of the performance of their models using only metadata. We attempt to fill the gap by proposing an AI-based framework to classify skin lesions solely based on metadata. The framework comprises collecting and analyzing metadata to identify relevant meta-features associated with

[1] Md Islam, Alba Garcia, John Gan and Haider Raza are with the School of Computer Science and Electronics Engineering, University of Essex, United Kingdom

[2] Gordon Wishart is with Check4Cancer Ltd., and Anglia Ruskin University, UK

[3] Joseph Walls is with Check4Cancer Ltd., and Fitzwilliam Hospital, UK

[4] Per Hall is with Check4Cancer Ltd., and Addenbrookes Hospital NHS Foundation Trust, UK

*This work was supported by Innovate UK Funding for Knowledge Partnership with Check4Cancer Ltd.

skin cancer progression followed by separating suspicious lesions from non-suspicious ones through applying AI models. This research work offers three major contributions:

1) Collection of 54,000 skin lesions metadata from 25,214 patients across a national network of private UK skin diagnostic clinics.
2) Identification of a subset of meta-features significantly related to the development of skin cancer.
3) Adaptation and optimization of five predictive AI models for skin lesion binary classification based solely on metadata.

This paper is structured as follows: Section II describes the metadata collection, feature selection, and classification model. Section III provides information on the results and discussion, and Section IV concludes the study.

II. Data and Method

A. Metadata Collection

In this study, we collected 54,000 skin lesions metadata from 25,214 patients, who attended private skin cancer diagnosis clinics between [2015-2022]. Ethical approval was received from the university Ethics Committee. The data collection summary is provided in Table I. All the features excluding lesion rating are used as input to the machine learning (ML) models to classify whether input features belong to suspicious or non-suspicious categories.

The meta-features listed in Table I are self-explanatory except for the lesion location feature that comprises seven values according to the anatomic location of the lesion such as i) Head and Neck, ii) Trunk waist up (front or back), iii) Groin/Buttocks/Genitals, iv) Hand, v) Foot, vi) Left/Right Leg (ankle up), and vii) Left/Right Arm (wrist up). The Williams score is calculated based on the method explained in the study [12] and summarized in Table II, where age, gender, sunburns, natural hair colour, the density of freckles on arms, number of moles, and prior non-melanoma history features were included to calculate the final Williams scores. The lesion score is calculated based on a weighted 7PCL as mentioned in the study [7] using the eq 1:

$$\text{Lesion Score} = 2\sum_{i=1}^{3} M_i + \sum_{j=1}^{4} N_i \tag{1}$$

where M is the set of major lesion features (change in size, shape, and colour) and N is set of minor features (inflammation, oozing, itching, and diameter $\geq$7 mm).

The 7PCL was first formulated by Mackie et al. [13], where they included seven lesion characteristics (change in size, shape, colour, inflammation, oozing, itching, and diameter $\geq$7 mm), each having a score of 1, to help prioritise pigmented skin lesions for urgent referral. Later Walter et al. [7] were able to confirm better results with a revised version, which separated lesion features into two groups:- i) major features (change in size, shape, and colour) each having a score of 2 and ii) minor features (inflammation, oozing, itching, and diameter $\geq$7 mm) with a score of 1. Consequently, lesions with scores $\geq$3 were sent for specialist opinion. Finally, the target variable, lesion rating, comprises two values (suspicious and non-suspicious) rated by our in-house skin cancer specialists. The experts classified pigmented lesions with atypical features in size, shape, color, or dermatoscopic appearance of melanoma as suspicious. Furthermore, skin lesions suspicious of either BCC, SCC, or potentially pre-malignant Actinic Keratoses were also rated as suspicious.

We have encoded all the non-numerical meta-features to convert them into categorical features using a one-hot encoding approach as summarized in Table III for an illustrative purpose. An effort was made to analyze the collected meta-features through explanatory data analysis (EDA). 95% of the skin lesions with a lesion score of zero belong to a non-suspicious class as illustrated using the bar plot in Fig. 1. Contrarily, more than 50% skin lesions with a lesion score of 10 fall in the suspicious category. Therefore, it can be inferred that the higher the lesion score, the higher the probability that the skin lesions belong to a suspicious group (p-value<0.01). For William's score between 56 and 61, around 60% cases belong to a suspicious category, whereas only around 7% cases belong to a suspicious group with Williams scores between 0–6 as summarized using bar plot in Fig. 1. The higher Williams score likely increases the chance of the skin lesion being suspicious as compared to a low Williams score.

Furthermore, we also analysed another potential meta feature - patient age - summarised using a probability density function in Fig. 2. We observed that the mean age of patients with a suspicious skin lesion is 52, which is significantly higher than the patients' mean age of 41 years with non-suspicious skin lesions (p-value<0.01).

B. Feature Selection

Three feature identification methods, i.e., wrapper, Shapley, and Pearson correlation, were utilized to select relevant features associated with skin cancer development. Consequently, the highly relevant attributes were identified and used during model training and evaluation.

The wrapper method [14] is an ML-based approach that works as a black-box function to evaluate subsets of features. The wrapper method produces a set of representative features and then uses an ML model to train and evaluate each subset. Based on the model's performance, the wrapper method identifies the best subset of features. The feature selection based on wrapper method is briefly summarized as follows:

- Subset generation: first, a subset of features is generated. This can be done in two ways- start with one feature and gradually add more, or start with all features and gradually remove them, or generate subsets of features randomly.
- Subset evaluation: after a subset of features has been generated, a model is trained on this subset of features, and the model's performance is evaluated, usually through cross-validation. The performance of the model gives an estimate of the quality of the features in the subset.

TABLE I

List of 25 meta-features: a total of 54,000 skin lesions metadata from 25,214 patients have been collected.

Meta-Feature	Description	Type	Range
Patient ID	Patients' ID anonymous	Alphanumeric	A-Z, a-z,0-9
Lesion ID	Lesions' ID anonymous	Alphanumeric	A-Z, a-z,0-9
Age	Patients' age in years	Numeric	0–92
Gender	Patients' gender at birth (M/F)	Categorical	0–1
Lesion Size	Change in size (yes/no)	Categorical	0–1
Lesion Age	Has it been present <6 months? (yes/no)	Categorical	0–1
Lesion Shape	Change in shape (yes/no)	Categorical	0–1
Lesion Colour	Change in colour (yes/no)	Categorical	0–1
Lesion >7mm	Is it 7mm or more? (yes/no)	Categorical	0–1
Lesion Inflamed	Is it Inflamed? (yes/no)	Categorical	0–1
Lesion Oozing	Is it Oozing? (yes/no)	Categorical	0–1
Lesion Pink	It is pink? (yes/no)	Categorical	0–1
Lesion Itch	Is it itchy? (yes/no)	Categorical	0–1
Lesion Location	Location on the body- Head Neck, Hand, Foot, Left/Right Leg, Left/Right Arm	Categorical	0–5
Williams Score	Williams score calculated based on [12]	Numeric	0–67
Prior Family History	Prior family history of skin cancer (yes/no)	Categorical	0–1
Williams Group	Williams group (<25 = average risk; 25+ = high risk)	Categorical	0–1
Hair	Natural hair color (black, red, blonde, brown)	Categorical	1–4
Sunburn	Number of sunburns (0, 1–4, 5–9, >10 burns)	Categorical	1–4
Mole	Number of moles (1, 2, 3 or more, none)	Categorical	1–4
Freckle	The density of freckles on arms (a few, several, a lot, none)	Categorical	1–4
Prior Melanoma	Any prior history of melanoma (yes/no)	Categorical	0–1
Prior Skin Cancer	Any prior history of skin cancer (yes/no)	Categorical	0–1
Lesion Score	7-point weighted checklist score based on [7]	Numeric	0–10
Lesion Rating	Target variable whether lesion is suspicious or non-suspicious	Categorical	0–1

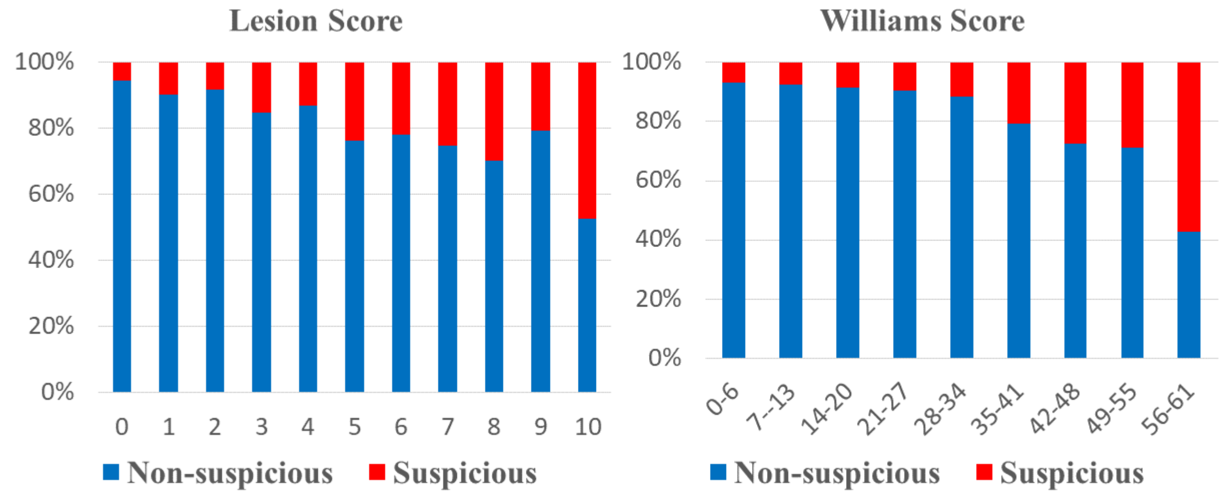

Fig. 1. Comparison of Lesion and Williams scores for suspicious and non-suspicious cases.

- Stopping criterion: this process is repeated, generating and evaluating different subsets of features, until some stopping criterion is met. This could be a certain number of subsets evaluated, a certain amount of time elapsed, or no improvement in model performance after a certain number of iterations.

Our proposed feed-forward wrapper technique adapts a random forest (RF) classifier as a base learner and iteratively estimates model performance for a subset of features. We used 10-fold cross-validation (CV) while evaluating the performance of the RF classifier as a feature selector to identify an optimal subset of meta-features with stopping criteria of no improvement in model performance after 1000 iterations.

ML interpretability is a topic of growing importance in ML. Interpretability is the ability to explain ML model decision making and nowadays researchers as well as users of ML models prefer to have some kind of explanation for informed decision-making. This is more important when dealing with real-world applications, particularly in healthcare applications. One such method frequently used for the interpretability of ML models is known as Shapely Additive Explanations (SHAP) [15]. This method is preferred over the traditional statistical-based methods such as filter because many of these methods can be inconsistent, which means that the most important features may not always be given the highest feature impor-

TABLE II
CALCULATION OF WILLIAMS SCORE BASED ON THE RISK FACTORS DESCRIBED IN THE STUDY [12].

Risk Factor	Category	Score
Gender	Female	0
	Male	7
Age	35-44	0
	45-54	5
	55-64	8
	65-74	11
Sunburn	None	0
	1-4	1
	5-9	4
	10 or more	7
Hair	Dark brown/Black	0
	Light brown	4
	Blond	5
	Red	8
Freckle	None	0
	Few	4
	Several	6
	A lot	10
Mole	None	0
	1	3
	2	5
	3 or more	11
Prior Skin Cancer	No	0
	Yes	13

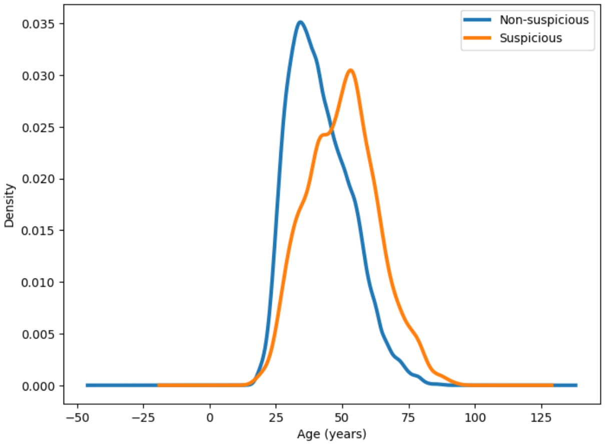

Fig. 2. Comparison of patient age distribution for suspicious and non-suspicious cases.

tance score. One example is that in the tree-based method as a wrapper which might give two equally important features different scores based on what level of splitting was done using the features. The features which split the model first might be given higher importance. This motivates us to use the SHAP method for feature selection for our use case. Shapley method assigns high scores to meta-features due to their performance in classifying instances correctly. Our adopted Shapley value-based feature selection method measures the marginal contribution of each feature when combined with other features.

Filter methods utilize univariate statistics to test whether there is a significant relationship between each input meta-feature to the target variable. The meta-features that provide the highest correlation values are the features that are kept for model development. One of the benefits of using the filter method is that, this method is not dependent on the ML models that we decide to develop. We adopted a filtering method known as Pearson's correlation coefficient to select relevant meta-features. The Pearson correlation [16] is a statistical approach that is frequently used in healthcare applications to measure the amount of linear correlation between an input X meta-feature and the output Y target variable. It ranges from +1 to -1, where 1 means there is a total positive correlation, and -1 means that there is a total negative correlation. Conversely, 0 means that there is no linear correlation. To calculate the Pearson correlation coefficient, we take the covariance of the input meta-feature X and output target variable Y and divide it by the product of the two variables' standard deviation, $i.e.$,

$$\rho_{X,Y} = \mathrm{corr}(X,\ Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

where cov and $corr$ denote the covariance and the correlation coefficient respectively, and σ_X andσ_Y are the standard deviation (SD) of the random variable X and Y, respectively.

C. Skin Lesion Classification Model

This study proposes an AI framework for suspicious vs non-suspicious skin lesion binary classification based solely on patient metadata. An overview of the proposed AI model is shown in Fig. 3. We adopted five ML models for skin lesion binary classification: suspicious vs non-suspicious categories based on patient metadata described in this section.

1) Naive Bayes (NB) classifier: It is a candid and compelling algorithm for the classification task based on the Bayes theorem. It predicts a class level's probability given a particular data record [17]. The class with the highest probability is considered as the predicted class for the given data tuple. NB classifiers assume that all attributes are conditionally independent of the given class label. The goal of this classifier is to learn a representative function from a given training labeled dataset. The conditional probability $p(Y|X)$ of target variable Y is calculated as follows:

$$p(Y \mid X) = \frac{p(Y)\ p(X \mid Y)}{p(X)} \tag{3}$$

where $p(Y)$ is the prior probability of a class Y, $p(X|Y)$ is the conditional probability of meta-feature given a particular class, and $p(X)$ is the evidence or probability of data X regardless of its target class (suspicious or non-suspicious).

2) Support vector machine (SVM): It is one of the most popular supervised ML approaches more frequently used for classification in various industries such as healthcare applications [18]. SVM finds a hyperplane to maximize the margin between the groups by utilizing the Lagrangian optimization technique [19]. One of the fundamental advantages of SVM is that if the data is linearly separable, then there is a unique global maximum value of the margin. In cases of non-linear distribution of the data, where a hyperplane cannot separate

TABLE III
METADATA CONVERSION USING ONE-HOT ENCODING APPROACH.

Raw metadata before pre-processing

Lesion Size	Lesion Shape	Gender	Lesion Rating
No	Yes	F	Green
Yes	Yes	M	Green
Yes	No	F	Green
Yes	Yes	F	Red
Yes	Yes	M	Red
Yes	No	M	Green
No	Yes	M	Red
Yes	Yes	M	Green
No	No	F	Red
Yes	Yes	F	Red

Metadata after pre-processing

Lesion Size	Lesion Shape	Gender	Lesion Rating
0	1	0	0
1	1	1	0
1	0	0	0
1	1	0	1
1	1	1	1
1	0	1	0
0	1	1	1
1	1	1	0
0	0	0	1
1	1	0	1

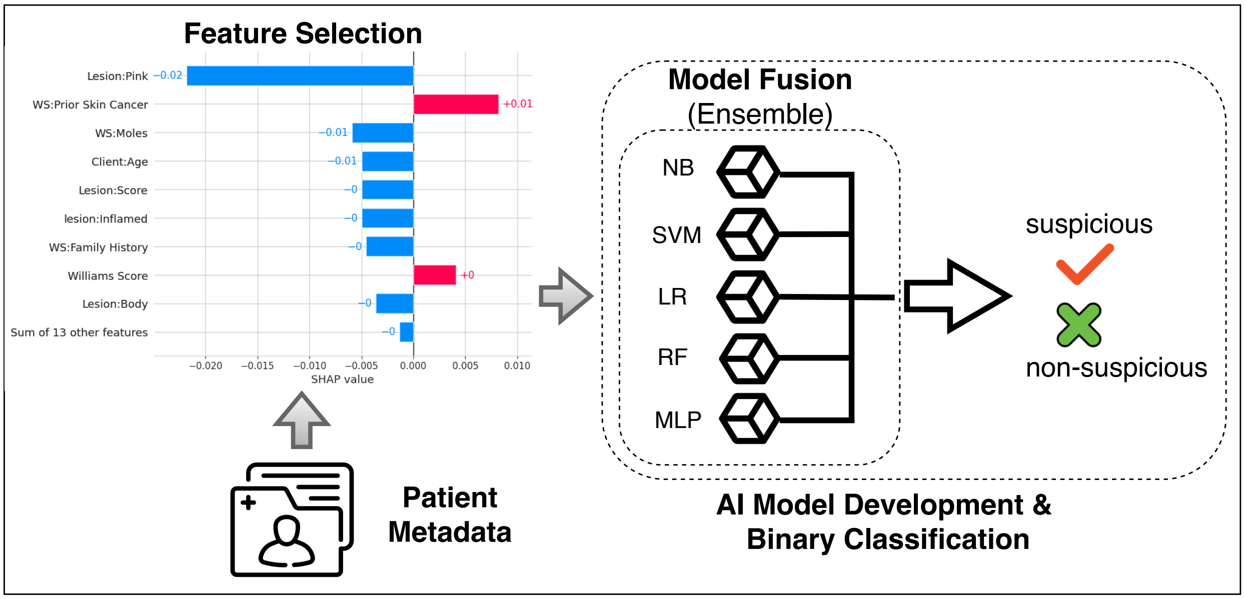

Fig. 3. The Proposed AI framework for skin lesion classification into suspicious and non-suspicious based on patient metadata.

the region, SVM uses a kernel function technique. The kernel function transforms the data into a higher dimensional feature space where the data's linear separation is possible.

3) Logistic regression (LR): It is a statistical method, where log-odds of the probability of an event are linear combinations of independent variables [20]. Although the model outputs the probability of an event, it is used in the classification task by applying a threshold. The logistic regression approach's outcome is binary, such as positive or 1 (suspicious) and negative or 0 (non-suspicious). Our adapted LR tries to develop a relationship (function) between the meta-feature and outcome variable by finding the best descriptive fitting model. Two different approaches were available for learning this function. A discriminating model learns the function directly to compute class posterior while a generative model learns the conditional class probability and class prior by applying Bayes rule [21]. We used a modified alternative to discriminative and generative models to merge probability altogether to learn the discriminative function, which directly maps input meta-feature to output target variable as follows:

$$p(Y|X) = \frac{exp(\beta_0 + \sum_{i=1}^{P} \beta_i X_i)}{1 + exp(\beta_0 + \sum_{i=1}^{P} \beta_i X_i)} \tag{4}$$

where $p(Y|X)$ is probability of a skin lesion being suspicious (Y=1) given meta-feature X, β_0 is the intercept, and β are the coefficient values, P is the total number of meta-features.

4) Multi-layer perceptron (MLP): It is one of the dominant predictive models used in machine learning. As the name 'neural' suggests, MLP is a brain-inspired system that tries to replicate the human brain [22]. MLP consists of an input and output layer and a hidden layer (in most cases) to transform input into some form that the next layer can use. An MLP is handy in finding a pattern or feature extraction from data that is considered complicated or laborious for a human. The success of neural network-based approaches such as MLP is due to a technique known as "backpropagation," which allows changing the weight of the hidden layer if there are any errors. The fundamental advantage of MLP is that it does not require in-depth knowledge about the relationship between input meta-feature and output target variables. Instead, it tries to recognize a pattern in the dataset and store those patterns as a weight for later use for the test cases. In our implementation, We have adopted an MLP with three hidden layers (32, 16, 8 neurons), rectified linear activation function (ReLU), and adaptive moment estimation (Adam) optimizer.

5) Random Forest: It employs an ensembling technique that generates multiple random trees and combines the outcome of a test sample based on majority voting or averaging [23]. During RF model development, trees are built upon a bootstrap sample of the data. RF adds more randomness in selecting a subset of predictors compared to a decision tree, where each node is split using the best variable selected based on a node splitting criterion - gini or entropy. This randomness in selecting features makes the RF classifier more accurate and robust compared to other classifiers such as SVM, discriminative analysis, and neural network [24]. In our adaptation, we optimized the RF model to find the best hyper-parameters (number of trees, 500, max depth, 40, splitting criterion, gini, bootstrap, true) for classifying skin lesions into suspicious and non-suspicious categories.

Furthermore, this study employed the ensembling technique of majority voting-based decision-making by combining NB, LR, SVM, RF, and MLP model outcomes. The final decision of a test sample is taken based on majority voting. In the stacking approach, we stacked NB, LR, SVM, and RF as feature extractors and MLP as meta-learners to classify input metadata into suspicious and non-suspicious classes.

D. Data Split and Evaluation Metrics

We split the metadata of 54,000 skin lesions into training (80%) and test (20%) sets. During the training, a 10-fold CV was used to build the models. The models were optimized to tune hyper-parameters and the best-performing models were selected based on their 10-CV results on training data. Consequently, the selected models were evaluated on test data. The following evaluation metrics were utilized to assess the performance of the developed AI framework:

$$\text{Sensitivity (Sen)} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{Specificity (Spc)} = \frac{TN}{FP+TN} \tag{6}$$

$$\text{Balanced Accuracy (Acc)} = \frac{Sen+Spc}{2} \tag{7}$$

where TP, TN, FP, FN refer to true positive (suspicious classified as suspicious), true negative (non-suspicious classified as non-suspicious), false positive (non-suspicious misclassified as suspicious), and false negative (suspicious misclassified as non-suspicious) instances, respectively.

III. Results and Discussion

In this section, the results of the feature selection are provided and analyzed. The performance of the developed AI framework is presented and benchmarked with those in the literature.

A. Feature Selection Results

The selected features using the wrapper, Shapley, and Pearson correlation methods are summarized in Table IV. The wrapper method ranks the number of moles, freckles, family history, lesion pink, lesion age, lesion score, and patient age as top features. The Shapley method prioritises prior skin cancer, Williams score, lesion location, lesion pink, and lesion score as top features. The Pearson correlation method ranks lesion pink, lesion inflamed, patient age, lesion oozing, lesion score as top features. The three most correlated features listed by all the feature selection methods are lesion pink, lesion score, and Williams score.

TABLE IV
Feature selection results for the wrapper, Shapley, and Pearson correlation methods.

Wrapper	Shaply	Pearson Correlation
1 Mole	1 Lesion Pink	1 Lesion Pink
2 Freckle	2 Prior Skin Cancer	2 Lesion Inflamed
3 Family History	3 Mole	3 Client Age
4 Lesion Age	4 Client Age	4 Lesion Oozing
5 Lesion Size	5 Lesion Score	5 Lesion Score
6 Lesion Pink	6 Lesion Inflamed	6 Lesion Itch
7 Lesion Body	7 Family History	7 Prior Skin Cancer
8 Client Age	8 Williams Score	8 Lesion Size
9 Williams Score	9 Lesion Body	9 Williams Score
10 Lesion Score	10 Lesion Size	10 Sunburn

TABLE V
The performance comparison of the RF model for different feature combinations (top 1-5, top-10, top-15, and all features).

Feature	**Acc**	**Sen**	**Spc**
Top1: Lesion Score	62.26%	57.87%	66.64%
Top2: Lesion Score Lesion Pink	68.12%	70.29%	67.95%
Top3: Lesion Score Lesion Pink Gender	68.85%	75.25%	62.44%
Top4: Lesion Score Lesion Pink Gender Williams Group	67.21%	79.02%	55.39%
Top5: Lesion Age Lesion Shape Lesion Pink Gender Williams Group	67.55%	81.23%	53.87%
Top10: Top5+ Lesion Body Lesion Score Prior Melanoma Age Freckle	68.26%	81.97%	54.55%
Top15: Top10+ Family History Mole Sunburn Prior Skin Cancer Lesion Oozing	70.53%	82.15%	58.91%
All Features from Table I	70.22%	83.16%	57.27%

Furthermore, we attempted to identify the top-performing features from all candidate features based on their individual

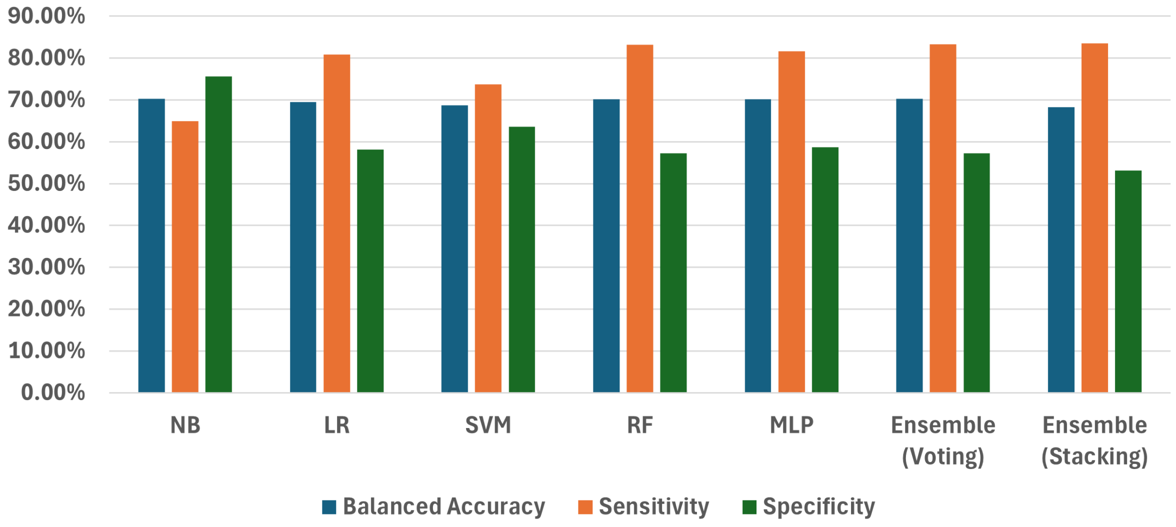

Fig. 4. The performance comparison of the employed ML models for skin lesions metadata classification.

contributions in classifying skin lesion metadata as suspicious or non-suspicious. The incremental performance gains for different top feature combinations (top 1–5, top–10, top–15, and all features) are summarized in Table V. We found lesion score as the top-1 contributing feature that alone achieved a balanced accuracy of 62.26%, a sensitivity of 57.87%, and a specificity of 66.64% using the RF model. Lesion pink and lesion score are found as the top-2 features that contributed about 12% sensitivity improvement. The RF model achieved a balanced accuracy of 67.55%, a sensitivity of 81.23%, and a specificity of 53.87% with the top–5 features (lesion age, lesion shape, lesion pink, patient gender, and Williams group). The RF model performed best for all the collected features listed in Table I with a balanced accuracy of 70.22%, a sensitivity of 83.16%, and a specificity of 57.27%.

Finding a subset of significantly correlated features has great potential to reduce workload during healthcare data collection. This can also significantly reduce data size and shorten model training time. Our research identified a set of top–5 meta-features from a pool of 25 features, which achieved a promising sensitivity of 81.97% (for all 25 features, sensitivity was 83.16%).

B. Skin Lesion Classification Results

The performance of the ML models is depicted in Fig. 4. The NB model attained a balanced accuracy of 70.31%, with a sensitivity of 64.95% and a specificity of 75.67%. Compared to the LR model, NB exhibited an improvement in sensitivity by approximately 16%, although specificity notably decreased to 58.13%. Subsequently, the RF model enhanced sensitivity to 83.16%. The MLP model yielded a balanced accuracy of 70.16%, with a sensitivity of 81.60% and a specificity of 58.72%. Employing a voting approach resulted in a sensitivity of 83.35%, comparable to that of the RF model. Notably, the stacking ensemble of NB, LR, MLP, and RF models achieved the highest sensitivity of 83.53%. Although RF, Voting, and Stacking models achieved similar performance, we selected RF models for benchmarking as RF as a single model offers less computational complexity, and training time compared to voting (five models) and stacking (five models) approaches.

In this study, sensitivity takes precedence over other evaluation metrics, notably specificity. This prioritization stems from the critical importance of accurately detecting suspicious lesions, particularly those associated with melanoma. Given the severe consequences of missing a suspicious lesion, our focus on sensitivity aims to minimize FN. Following discussions with our in-house skin cancer experts, we opted to trade off specificity scores to enhance the detection of suspicious cases.

TABLE VI
THE PERFORMANCE COMPARISON OF THE RF MODEL FOR DIFFERENT COMBINATIONS OF THE OUTCOME PROBABILITY THRESHOLD.

Threshold	Balanced Accuracy	Sensitivity	Specificity
0.01	59.63%	94.94%	24.32%
0.02	64.04%	92.00%	36.08%
0.03	67.01%	88.96%	45.06%
0.04	68.72%	85.92%	51.52%
0.05	**70.22%**	**83.16%**	**57.27%**
0.06	71.33%	80.59%	62.07%
0.07	73.17%	77.18%	69.15%
0.08	73.13%	74.33%	71.93%
0.09	73.48%	72.49%	74.46%
0.10	73.29%	69.83%	76.76%

There was a compromise between sensitivity and specificity scores. As we aimed for high sensitivity, the specificity scores were affected as highlighted in Table VI. In our experiment, the ML models used a default probability threshold value of 0.50 for classifying input data into output classes. We have tuned this probability threshold value between 0 and 1 with an interval of 0.01. The RF model outperformed in terms of sensitivity (83.16%) with a threshold value of 0.05 as compared to a default threshold value of 0.50 (sensitivity 77.15%). Although the sensitivity increased with a 0.05 threshold value,

TABLE VII
THE PERFORMANCE BENCHMARKING OF OUR APPROACH ALONG WITH THE 7PCL AND WILLIAMS METHOD.

Method	7PCL Method [7]	Williams Method [12]	Our Approach
Feature	1.Lesion Size 2.Lesion Color 3.Lesion Shape 4.Lesion >7mm 5.Lesion Inflamed 6.Lesion Oozing 7.Lesion Itch	1.Patient Gender 2.Patient Age 3.Sunburn 4.Hair Color 5.Mole 6.Freckle 7.Prior Skin Cancer	All the meta-features listed in Table I
AI Model Performance	Balanced Accuracy: 64.58% Sensitivity: 68.09% Specificity: 61.07%	Balanced Accuracy: 64.01% Sensitivity: 66.32% Specificity: 61.71%	Balanced Accuracy: **70.22%** Sensitivity: **83.16%** Specificity: 57.27%

other evaluation metrics such as specificity dropped to 57.27%.

We have benchmarked our approach with the 7PCL [7] and Williams method [12] and the comparative performance gain is highlighted in Table VII. Our employed RF model outperformed the 7PCL and Williams methods in both balanced accuracy (p-value<0.01) and sensitivity (p-value <0.01). The 7PCL method achieved 64.58% balanced accuracy, 68.09% sensitivity, and 61.07% specificity when evaluated using our dataset. Although the Williams method identifies a person at risk of developing skin cancer, we utilized Williams features to investigate whether a person's lesion can be classified as suspicious or non-suspicious based on the Williams features. The Williams method displayed a similar performance as the 7PCL method with a balanced accuracy of 64.01%, a sensitivity of 66.32%, and a specificity of 61.71%. Conversely, our approach utilizing the RF model along with all the meta-features listed in Table I achieved a significant performance gain over the 7PCL and Williams methods. The RF model displayed 70.22% balanced accuracy, 83.16% sensitivity, and 57.27% specificity. We prioritize sensitivity over specificity as missing a suspicious case is more severe than missing a non-suspicious case. Therefore, although we achieved a lower specificity as compared to the 7PCL and Williams method, our sensitivity is outperformed by 15%.

Moreover, to compare our approach with ML-based methods, we conducted a literature review. We noted a study by Pacheco et al. [9], which incorporated patient metadata such as age, gender, lesion location, bleeding, and pain alongside skin images. They reported a 7% enhancement in performance attributed to the inclusion of metadata. However, they did not separate the impact of metadata alone on skin cancer detection. Another study by Ha et al. [25], the winner of the Kaggle 2020 melanoma challenge, integrated patient metadata such as age, gender, and lesion location. Interestingly, they found that augmenting images with metadata did not improve model performance. Previous research predominantly relied on image data alone, with limited exploration of patient metadata for lesion classification in skin cancer detection. Hence, we developed an AI framework exclusively leveraging metadata. Our findings indicate that this framework effectively discriminates between suspicious and non-suspicious skin lesions with high sensitivity. This approach has the potential to complement existing skin cancer assessment methods when used alongside image data. In the future, patients categorized as non-suspicious based on both metadata and images could potentially avoid referrals to specialist clinics. Additionally, metadata classification could serve as a decision support tool for telemedicine reporters when lesion classification remains uncertain post-image analysis alone. This approach holds promise for reducing the number of referrals for potential biopsies and alleviating waiting times for skin cancer diagnosis.

IV. CONCLUSION

The integration of AI methodologies in skin lesion classification, focusing solely on metadata, holds significant promise in streamlining and expediting the detection of suspicious lesions. By potentially reducing patient referrals for biopsy procedures, this approach has the potential to mitigate waiting times for skin cancer diagnosis and treatment, ultimately enhancing patient outcomes. In our study, we developed an AI framework exclusively leveraging patient metadata for skin lesion classification, achieving a sensitivity of 83.53%. Additionally, our research contributed to the acquisition of high-quality data and the identification of a subset of meta-features highly pertinent to skin cancer development. Future efforts may involve amalgamating these highly correlated meta-features with image modalities and employing computer vision models, which could potentially augment the performance of the classification model.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA: a cancer journal for clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[2] M. Pacifico, R. Pearl, and R. Grover, "The uk government two-week rule and its impact on melanoma prognosis: an evidence-based study," *The Annals of The Royal College of Surgeons of England*, vol. 89, no. 6, pp. 609–615, 2007.

[3] L. Smith, N. Sansom, S. Hemphill, S. H. Bradley, B. Shinkins, P. Wheatstone, W. Hamilton, and R. D. Neal, "Trends and variation in urgent referrals for suspected cancer 2009/2010–2019/2020," *British Journal of General Practice*, vol. 72, no. 714, pp. 34–37, 2022.

[4] C. Garbe, T. Amaral, K. Peris, A. Hauschild, P. Arenberger, N. Basset-Seguin, L. Bastholt, V. Bataille,

V. Del Marmol, B. Dréno *et al.*, "European consensus-based interdisciplinary guideline for melanoma. part 1: Diagnostics: Update 2022," *European Journal of Cancer*, vol. 170, pp. 236–255, 2022.

[5] M. Dildar, S. Akram, M. Irfan, H. U. Khan, M. Ramzan, A. R. Mahmood, S. A. Alsaiari, A. H. M. Saeed, M. O. Alraddadi, and M. H. Mahnashi, "Skin cancer detection: a review using deep learning techniques," *International journal of environmental research and public health*, vol. 18, no. 10, p. 5479, 2021.

[6] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Transactions on Medical Imaging*, vol. 20, no. 3, pp. 233–239, 2001.

[7] F. M. Walter, A. T. Prevost, J. Vasconcelos, P. N. Hall, N. P. Burrows, H. C. Morris, A. L. Kinmonth, and J. D. Emery, "Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study," *British Journal of General Practice*, vol. 63, no. 610, pp. e345–e353, 2013.

[8] I. Papachristou and N. Bosanquet, "Improving the prevention and diagnosis of melanoma on a national scale: A comparative study of performance in the united kingdom and australia," *Journal of Public Health Policy*, vol. 41, pp. 28–38, 2020.

[9] A. G. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," *Computers in biology and medicine*, vol. 116, p. 103545, 2020.

[10] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Experimental dermatology*, vol. 27, no. 11, pp. 1261–1267, 2018.

[11] D. Roffman, G. Hart, M. Girardi, C. J. Ko, and J. Deng, "Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network," *Scientific reports*, vol. 8, no. 1, p. 1701, 2018.

[12] L. H. Williams, A. R. Shors, W. E. Barlow, C. Solomon, and E. White, "Identifying persons at highest risk of melanoma using self-assessed risk factors," *Journal of clinical & experimental dermatology research*, vol. 2, no. 6, 2011.

[13] R. M. MacKie, *An illustrated guide to the recognition of early malignant melanoma*. University of Glasgow, 1986.

[14] I. M. El-Hasnony, S. I. Barakat, M. Elhoseny, and R. R. Mostafa, "Improved feature selection model for big data analytics," *IEEE Access*, vol. 8, pp. 66 989–67 004, 2020.

[15] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *Ieee Access*, vol. 9, pp. 144 352–144 360, 2021.

[16] D. Risqiwati, A. D. Wibawa, E. S. Pane, W. R. Islamiyah, A. E. Tyas, and M. H. Purnomo, "Feature selection for eeg-based fatigue analysis using pearson correlation," in *2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, 2020, pp. 164–169.

[17] K. P. Murphy *et al.*, "Naive bayes classifiers," *University of British Columbia*, vol. 18, 2006.

[18] E. Byvatov and G. Schneider, "Support vector machine applications in bioinformatics." *Applied bioinformatics*, vol. 2, no. 2, pp. 67–77, 2003.

[19] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[20] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 215–242, 1958.

[21] F. E. Harrell, "Ordinal logistic regression," in *Regression modeling strategies*. Springer, 2015, pp. 311–325.

[22] T. M. Mitchell, "Artificial neural networks," *Machine learning*, vol. 45, pp. 81–127, 1997.

[23] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[24] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 313–325.

[25] Q. Ha, B. Liu, and F. Liu, "Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge," *arXiv preprint arXiv:2010.05351*, 2020.