Journal of Robotics and Control (JRC)
Volume 6, Issue 1, 2025
ISSN: 2715-5072, DOI: 10.18196/jrc.v6i1.23302

AI-Driven Classification of Children's Drawings for Pediatric Psychological Evaluation: An Ensemble Deep Learning Approach

Ali Ibrahim Khlaif [1*], Mohamed Saber Naceur [2], Monji Kherallahr [3]
[1] National School of Engineers of Sfax, University of Sfax, Sfax, Tunisia
[2] University of Carthage, Tunis, Tunisia
[3] Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia
Email: [1] alialalusi00@gmail.com, [2] naceurs@yahoo.com, [3] monji.kheralleh@fss.usf.tn
*Corresponding Author

***Abstract*—In the wake of contemporary challenges such as the COVID-19 pandemic, understanding children's mental health through non-verbal forms like drawing has become paramount. This study enhances pediatric psychological assessments by employing an ensemble of deep learning models to interpret children's drawings, aiming for early detection of psychological states. Traditional drawing analysis methods are often subjective, variable and time consuming. To ddress these limitations, we developed an ensemble model that combines the strengths of VGG16, VGG19, and MobileNet architectures using a hard voting mechanism. This approach reduces bias and enhances reliability by integrating the unique capabilities of each model. Our methodology involved rigorous data collection through a custom Android application, followed by exploratory data analysis, data preprocessing, and comprehensive model valuation. The ensemble model was trained and validated on the diverse Kids' Hand Movement Dataset (KHMD), demonstrating superior accuracy and robustness in classifying drawings that indicate various psychological conditions. It significantly outperformed individual models, achieving a 99% accuracy rate. These findings underscore the potential of advanced machine learning techniques in providing more accurate and bias-free insights into children's psychological health, suggesting that ensemble learning can greatly improve the precision of pediatric psychological evaluations. Future work will explore expanding the dataset and employing more sophisticated ensemble methods to further enhance diagnostic accuracy.**

Keywords—Pediatric Psychological Assessment; Ensemble Deep Learning; Children's Drawings Analysis; Mental Health Detection; Art Therapy in Psychology.

I. INTRODUCTION

Growing complexities in societal pressures and competitive environments have raised significant concerns about mental well-being, particularly regarding children's mental health [1]. The COVID-19 pandemic has exacerbated these concerns, significantly impacting children by isolating them socially and increasing stress over their health and safety [2]. The early detection and treatment of mental health issues in children, including conditions such as depression, anxiety, ADHD, trauma, and developmental disabilities, are crucial; undetected, these issues can adversely affect long-term well-being into adulthood. Given the increasing prevalence of mental health issues among children, effective assessment tools are essential for early detection and intervention. Art therapy, particularly through drawing, has emerged as a vital tool in pediatric psychological evaluations due to its accessibility and expressiveness. Children often express their feelings and thought. through art more freely than verbally, making art therapy an effective means of accessing their subconscious. This is particularly beneficial for children who struggle to articulate their emotions, as drawing can help unbind suppressed thoughts and feelings. Therefore, art therapy has already become rather widespread due to its efficiency in unbinding suppressed thoughts and emotions [3]. Various projective drawing tests, such as the Bender Gestalt Test and the House-Tree-Person test, operate on the premise that drawings can reflect personality and emotional states, providing valuable insights into cognitive and personal development. These methods relevance is derived from their ability to provide insight into multiple psychological domains – cognitive and personal development, social connection – using a medium familiar and engaging to a child. Nonetheless, given the subjectivity and time-consuming nature of traditional drawing analysis, applying artificial intelligence for interpretation presents a promising alternative. The absence of an automated model for detecting mental health issues in children through their drawings highlights a critical gap in current methodologies.

Deep learning techniques can automate the analysis, offering bias-free psychological insights into children's emotional states [4]. The application of artificial intelligence in interpreting drawing tests presents a viable alternative to traditional methods. Research indicates that AI can effectively identify a diverse array of personality traits in children based on their drawings [5]. Machine learning techniques, such as Support Vector Machines (SVMs) and k-Nearest Neighbors (kNN), are supervised methods that have been successfully applied to classify features in hand-drawn images [6]-[11]. Additionally, deep learning—a subset of AI—enables the development of models that autonomously interpret drawings. These models are particularly valuable for providing unbiased psychological insights into a child's emotional state. Since its inception, deep learning has been employed in various methods to analyze hand-drawn images [12]-[19]. Convolutional Neural Networks (CNNs) excel at

extracting patterns from images, enabling them to identify hierarchical features directly from the visual data of hand-drawn items [16][20][21]. Transfer Learning utilizes pre-trained models, originally developed on extensive datasets, for hand-drawn recognition tasks. These models are then adapted or fine-tuned to meet specific needs [22]. Ensemble Methods Combining the predictions of multiple base classifiers through techniques such as bagging and boosting significantly enhances the overall performance of classification tasks in hand-drawn analyses [23]. This study advances the field by rigorously evaluating and comparing the performance of deep learning models—namely VGG16, VGG19 [24], and MobileNet—alongside an ensemble hard voting model [25][26]. Our approach harnesses the distinct strengths of each model while mitigating their individual weaknesses. By integrating diverse perspectives through hard voting, the ensemble model enhances prediction reliability and reduces the risk of overfitting.

The structure of this paper is as follows: Section 2 discusses related works, contextualizing our study within existing research on children's drawing analysis and deep learning applications. Section 3 details our methodology, including data collection, exploratory data analysis, data preprocessing, modeling techniques, and evaluation measures. Section 4 presents the experiments and results from each neural network model. Section 5 discusses the efficacy of ensemble methods based on the results. Section 6 concludes the paper, summarizing findings and suggesting future research avenues.

II. RELATED WORKS

The understanding of human behavior through handwriting and drawing analysis has witnessed significant advancements through the integration of deep learning techniques.

The research paper by Pysal et al. [27] investigates children's drawing strategies for creating seriation objects, focusing on the sequencing and order of strokes. Previous studies have identified six logical structures underlying these strategies. However, traditional evaluation methods relying on human observation are prone to inaccuracies. To address this, the study extends the research to touch screen drawings and employs a novel deep learning hybrid model (FLSTM) to classify drawing strategies. An application was developed for the drawing task, involving 32 children aged 5 to 12, resulting in 420 drawings. Comparative analysis with existing models (LSTM, CNN, Fuzzy-CNN) showed that the FLSTM model outperformed others with precision, recall, and F1 scores of 89.1%, 88.6%, and 88.6% respectively. The study demonstrates how deep learning facilitates understanding human psychological behavior through children's drawing analysis.

Ahmadsaraei et al. [28] explores the interpretation of psychological data through children's drawings, noting variations in behavior across different ages. The study compares various deep learning methods on children's data. After introducing the dataset and conducting statistical analysis, two methods, modified YOLO V5 (MYOLO V5) and modified ResNet 50 (MResNet 50), are compared on the OBGET dataset. The classification accuracy of these methods is evaluated on 386 Original Bender Gestalt Drawing Test samples for children. Preprocessing and semi-automatic labeling are conducted to prepare for comparison. Both MYOLO V5 and MResNet 50 achieve acceptable pattern detection accuracy, but MYOLO V5 demonstrates higher accuracy when applied to the collected dataset.

In their study, Shi et al. [29] address autism spectrum disorder (ASD) as a significant mental health concern globally. They highlight the lack of systematic comparison between paintings by ASD individuals and typically developed (TD) children. To address this gap, they create an ASD painting database comprising 478 paintings by ASD individuals and 490 by TD children. Through subjective and objective analysis, they identify key characteristics such as structuring logic, facial depiction, repetitive structures, composition location, and edge completeness in ASD paintings. Additionally, they develop a classifier based on these features to distinguish between ASD and TD painters, showing promising accuracy as a potential screening tool for ASD. Their work offers insights into understanding the distinctive aspects of autistic children's expressions through their artwork, with plans to release the database to the public.

Kamran et al. [30] address the challenge of diagnosing Parkinson's disease (PD) in its early stages. They highlight the importance of early diagnosis for effective symptom management, as PD currently lacks a cure. Leveraging handwriting samples from multiple PD datasets, they propose a method for early PD diagnosis using deep transfer learning-based algorithms. By combining HandPD, NewHandPD, and Parkinson's Drawing datasets, they achieve remarkable accuracy of 99.22% in identifying PD cases. Their approach demonstrates superiority over existing methods, offering promising potential for early PD detection and management.

Zhang et al. [31] focus on the importance of early detection in Parkinson's disease (PD) and the role handwriting analysis can play in diagnosis, given that hand tremors and handwriting difficulties are early motor symptoms. They propose a deep learning-based system for assessing patients' handwriting drawings as a means of early-stage PD diagnosis. Utilizing two datasets, HandPD and NewHandPD, consisting of hand-drawn spirals and meanders from PD patients and healthy individuals, they employ various deep learning models for classification. EfficientNet-B1 emerges as the most effective, achieving high precision, sensitivity, and accuracy in classifying patients' meander traced graphics. Additionally, they develop a userfriendly Windows application and a Python Web API based on Flask for the assessment system, enhancing its accessibility for screening tests and aiding healthcare professionals in Parkinson's disease diagnosis.

Nadeem et al. [32] examine the possibility of using handwriting analysis to detect emotions. According to the authors, a person's emotional state at the time of writing can be expressed in his or her handwriting. Previously, emotional states were recognized through visual, auditory analysis, and other nonverbal methods. In this paper, it is proposed to do this by handwriting. It is a relatively new area of research as earlier studies were limited to identifying such basic emotions as anxiety, stress, and depression. The main

purpose of this experiment was to identify the emotional state of the author, who can characterize a person as distressed/sad and suggest consulting a mental health professional. In addition, the authors evaluated how handwriting changes in different emotional states.

Elngar et al. [33] conducted a study devoted to handwriting analysis, a traditional method used for evaluating personality traits. However, now it has been adapted to find correlations between features of handwriting and one's personality traits of the Big Five Personality Traits. The authors suggested a dataset where samples of handwriting were checked against personality traits. They proposed a new algorithm, which includes two neural networks: Artificial Neural Network and Convolutional Neural Network. The performance of this approach is evaluated by comparing it with "traditional machine learning classifiers that evaluate key classification metrics". The tests demonstrated a high level of improvement over the traditional approach from Pearson and other authors.

Ghosh et al. [34] described an automatic graphology-based handwriting analysis method for analyzing human behavioral characteristics from structural components of handwritten English lower-case alphabets. The proposed system is novel in that it does not require human interpretation and instead analyzes individual characters based on loop, slant, and stroke. The proposed system is easily accessible online, where the end-users input the characters and the system analysis their personal and social characteristics. The feedback from the users of different genders and age groups was analyzed to score 86.70%, with 5300 different responses collected from the end-users of the system.

Mekhaznia et al. [35] analyzed the personality trait identification from people's handwriting, which is highly applied in forensic science, recruitment, and healthcare. Despite the uniqueness of each human's handwriting characteristics, personality traits recognition presents an immense problem. This paper developed an architecture to extract textural features of handwriting applications and apply artificial neural networks on the TxPI-u database. The proposed architecture has shown an impressive recognition rate compared with existing functional ontologies in identifying the personality traits of individuals.

Saraswal et al. [36] presented an automated system capable of personality prediction from an individual's handwriting to save laborious work for a graphologist. The authors introduced the K-nearest neighbor algorithm to the system for analyzing handwriting characteristics such as letter size, word spacing, slant, pen pressure, and baseline. The research was done to bring an effect to this growing interest in biometric and behavior-based personality prediction system wide.

In another work, Hamdi et al. [37] proposed the integration of BEM with a multi-stage DL-RNN-based approach that uses BLSTM. The goal of this method was to optimize online handwriting trajectory modeling for mobile and other fast devices in computation time, and space, by learning an end-to-end description system. The model covered preprocessing, segmentation, and trajectory approximation using the neural computation sequences in velocity and geometric profiles. The effectiveness of their methodologies was measured using mean absolute error and root mean square error on 2 datasets: LECTURE MULTI-CHANNEL Architecture and CALLIGRAPHY single-channel architecture, resulting in RMSE values 3.75% and 5.26% and MAE values 1.69% and 2.75%, respectively.

A novel multi-stage DL-based algorithm for multilingual online handwriting recognition has been proposed by Hamdi et al. [38]. It integrates a hybrid deep Bidirectional Long Short Term Memory (DBLSTM) and SVM networks. The method consists of initial preprocessing of the script, segmentation of Online Handwriting Trajectories (SOHTs), extraction of two types of feature vectors through the Beta-Elliptic Model (BEM) and CNN, classification of these vectors into subgroups through DBLSTM networks in a supervised mode, obtained via the Unsupervised Fuzzy K-Means algorithm for the online and offline branches, and combining these models by a trained SVM engine to increase the discrimination power. Extensive testing on three diverse datasets proved the effectiveness and synergy of the separate modules as well as the boost introduced by the overall fusion.

When experimenting with the classification of children's hand drawings as normal or not normal, we opted for an ensemble learning model that utilizes the hard voting technique. The choice was influenced by the unique and complex data variety within the organized dataset, which included children's drawings classified as either normal or not normal. In contrast to previously highlighted methodologies utilizing various deep learning architectures such as VGG16, VGG19, and MobileNet, this study presents a unique multiple models ensemble learning approach. The variability across possible classification parameter traits requires various classification frameworks. This framework is distinct from the previous relatively single-model various research methodologies since it unifies all the classification models toward the individual variances in the children's drawings and their potential classification, which took into consideration the psychometric distinctions between a normal or not normal drawing. The ensemble method using the hard voting mechanism effectively integrates all the classifiers' decision-making processes into a single optimal decision-making process, which is vital due to the unpredictable classification per image among markers. The hard voting model integrates the diverse model output to aid in dumping individual decision-making processes and providing TSU to a final decision-making process. The uniqueness in differences such as line thickness, the simplicity or complexity of the shape used, the overall composition, and the clear distinct object and color objects make the integration of the models effective. The unique classification in the TSU implementation that eliminates individual level misclassification offers a considerable significance in the overall implementation of utilization.

The comparative analysis of our method with related works is summarized in Table I. This table highlights the various datasets, age groups, types of data collected, models used, and key results from each study. Our approach, utilizing the Kids' Hand Movement Dataset (KHMD), involves a

combination of VGG16, VGG19, MobileNet to create an ensemble voting.

III. METHODOLOGY

The Fig. 1 illustrates our proposed methodology for analyzing children's drawings to assess psychological and developmental traits. Our study employs the Kids' Hand Movement Dataset (KHMD), which uses a specially designed Android application to capture the drawing movements of children aged 5 to 10 years. This application records children drawing shapes like ellipses and spirals, chosen to reveal insights into their motor skills and cognitive abilities. We standardize the recording environment by ensuring all children are seated at a consistent height and use a tablet for their drawings. For analysis, we visually explore the drawing patterns through Exploratory Data Analysis (EDA), which helps identify characteristics that categorize drawings as "normal" or "not normal". This visualization facilitates further analysis using deep learning models. In preprocessing, we normalize the image data and adjust dimensions to prepare for modeling, ensuring our models can generalize well from our training data. We use a custom-built Convolutional Neural Network that includes advanced architectures like VGG16, VGG19, and MobileNet. This model setup enhances the prediction accuracy by leveraging the combined strengths of these architectures through a hard voting ensemble method. Our methodology systematically progresses from data collection to evaluation, using metrics to assess each model's performance thoroughly. This structured approach ensures the study's integrity and the usefulness of our findings in enhancing psychological and developmental assessments for children.

A. Datset Collection

In our study focusing on analyzing children's hand drawing movements, we carefully crafted an Android application specifically for drawing particular geometric shapes—ellipses and spirals. Our target group was children between the ages of 5 to 10 years. To ensure a controlled and consistent data gathering environment, we made sure each child participant sat comfortably on a chair set at a height of 45 cm and used a table that was 80 cm high for their drawing activities on a tablet.

The protocol was structured to naturally capture the children's drawing behaviors. To initiate, an experimenter demonstrated the task to the participants, showing them how to trace the figures on the touchscreen using their finger. Following the demonstration, children were given a three-minute practice session to familiarize themselves with the task and the application interface. This preparation was crucial to ensure that the collected data reflected their best understanding and execution of the task.

During the actual data recording, children were asked to continuously trace the outlined figures for 30 seconds at their preferred speed, without specific instructions on movement velocity, emphasizing the spontaneity and comfort of their natural drawing pace (Fig. 2). The application interface, as shown in the Fig. 3, featured a simple 'Save' button, which allowed for the storage of the hand movements in both text (.txt) and image (.png) formats, thus facilitating subsequent data analysis.

TABLE I. COMPARATIVE STUDY OF RELATED WORKS

Study	Dataset Name	Age Group	Data Collected	Models Used	Key Results
Pysal et al. [27]	Seriation Objects Drawing Dataset	5-12 years	Sequencing and order of strokes	FLSTM, LSTM, CNN, Fuzzy-CNN	FLSTM model achieved precision, recall, and F1 scores of 89.1%, 88.6%, and 88.6% respectively
Ahmadsaraei et al. [28]	OBGET Dataset	Various ages	Psychological data through drawings	MYOLO V5, Mres-Net 50	MYOLO V5 demonstrated higher accuracy in pattern detection
Shi et al. [29]	ASD Painting Database	ASD and TD children	Paintings by ASD and TD children	Custom based features on classifier identified	Key characteristics identified for ASD paintings, promising accuracy as a screening tool for ASD
Kamran et al. [30]	HandPD, NewHandPD, Parkinson's Drawing datasets	Various ages	Handwriting samples	Deep transfer learning algorithms	Achieved 99.22% accuracy in identifying PD cases
Zhang et al. [31]	HandPD, NewHandPD	PD patients and healthy individuals	Hand-drawn spirals and meanders	EfficientNet-B1, Other deep learning models	High precision, sensitivity, and accuracy in classifying meander traced graphics
Nadeem et al. [32]	-	Various ages	Handwriting samples	-	Identified emotional states through handwriting, suggesting mental health consultations
Elngar et al. [33]	Personality Detection Dataset (PDD)	Various ages	Handwriting samples	ANN, CNN	High improvement over traditional methods in evaluating personality traits
Ghosh et al. [34]	Graphology-based Analysis	Various ages	Handwritten English lowercase alphabets	Novel system based on structural components	Scored 86.70% with user feedback
Mekhaznia et al. [35]	TxPI-u Database	Various ages	Handwriting samples	ANN	Impressive recognition rate in identifying personality traits
Saraswal et al. [36]	Automated Personality prediction from handwriting characteristics	Various ages	Handwriting characteristics	K-nearest neighbor algorithm	Automated personality prediction from handwriting characteristics
Hamdi et al. [37]	LECTURE MULTI-CHANNEL, CALLIGRAPHY single-channel	Various ages	Online handwriting trajectory	DL-RNN, BLSTM	Effective trajectory modeling with RMSE 3.75% and 5.26%, MAE 1.69% and 2.75%
Hamdi et al. [38]	Multiple datasets	Various ages	Multilingual online handwriting	DBLSTM, SVM, CNN	Proven effectiveness and synergy of modules with improved discrimination power

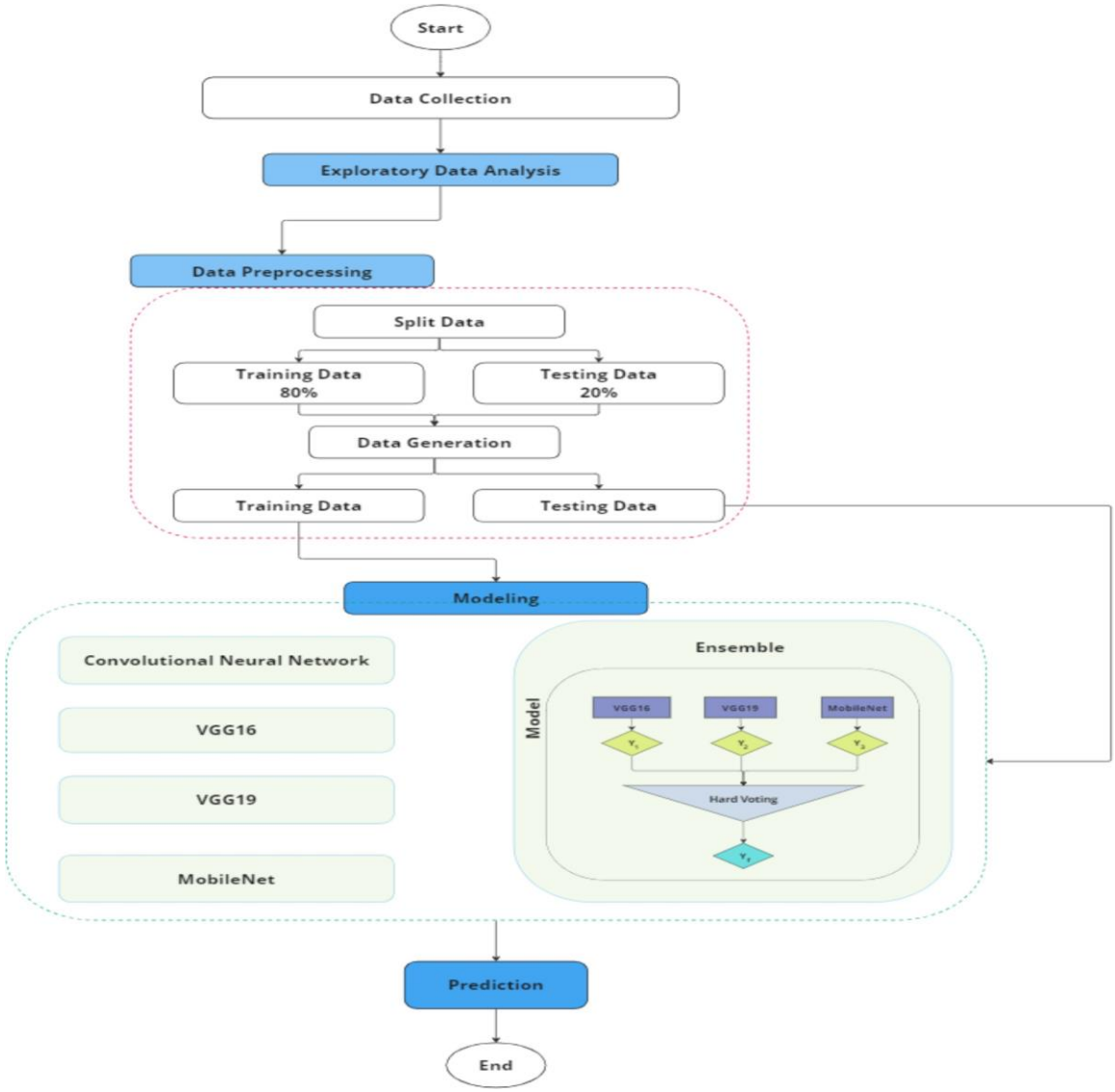

Fig. 1. Proposed approach

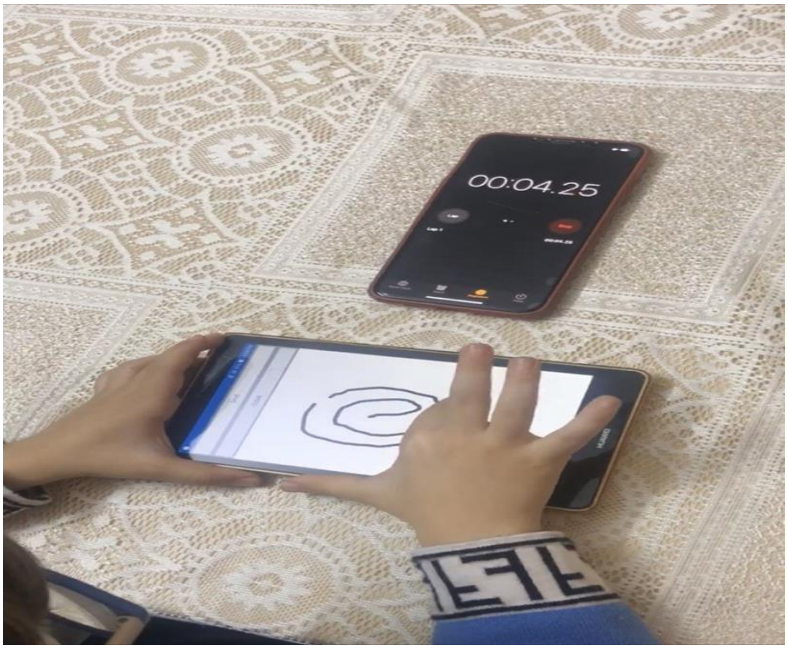

Fig. 2. Child participant tracing figures on tablet

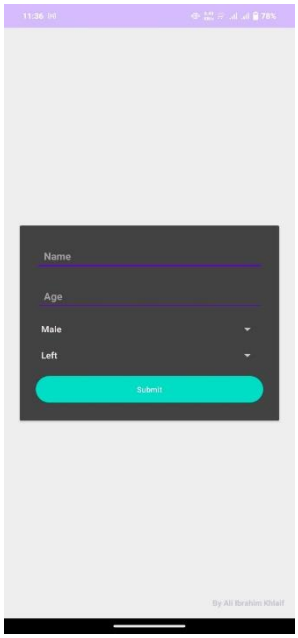

Fig. 3. Application interface

The design of this data collection approach was critical in ensuring the authenticity and consistency of the drawing data, which are paramount for the subsequent analysis using deep learning techniques. This methodological rigor aids in creating a robust dataset that accurately represents the motor skills and cognitive abilities reflected in the children's drawings, crucial for our study's aim to analyze these movements for psychological and developmental insights.

We collected an overall of 250 normal and 250 not normal images, ensuring a balanced Kids' Hand Movement Dataset (KHMD) for our analytical models. This comprehensive collection of drawing data forms the foundation for our further investigations into classifying and understanding the nuances of children's drawing behavior, as presents in Fig. 4.

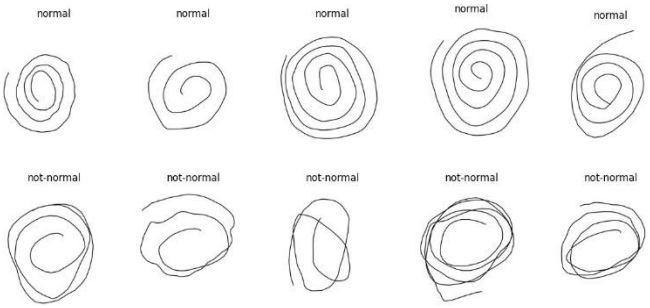

Fig. 4. Examples of hand-drawn figures

In our study, we sought participant diversity in socioeconomic, educational, and other demographic areas, recognizing their impact on children's drawing skills and cognitive development. We selected ellipses and spirals for tasks to evaluate fine motor control and cognitive processing, markers of neurological development. We also acknowledge the limitations in generalizability due to the controlled setting of chair and table heights, which may not fully capture natural drawing behaviors. These considerations are factored into our analysis to ensure the findings are contextualized correctly, enhancing the study's depth and relevance in assessing the relationship between children's motor skills and cognitive growth.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical initial phase in the data analysis process where researchers analyze, summarize, and visualize data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations. It allows researchers to understand the data's underlying structure, gain insights that are not immediately obvious, and inform subsequent analysis strategies and modeling [39].

Fig. 4 illustrates examples of hand-drawn figures from our Kids' Hand Movement Dataset (KHMD) categorized as "normal" and "not-normal" from the dataset used in our study. These images show variations in drawing styles and qualities that might indicate differences in motor skills, cognitive abilities, or psychological states among children. The "normal" drawings display relatively uniform and consistent spiral and elliptical shapes, indicating a level of precision and control in the drawing process. Conversely, the "not-normal" drawings are irregular, with more erratic lines and less symmetry, suggesting potential variations in the psychological or developmental states of the children. These visual distinctions form the basis for using machine learning models to classify and analyze these drawings systematically.

The bar chart shown in Fig. 6 represents the class distribution of samples in our dataset, specifically categorized into "normal" and "not-normal" classes based on children's hand drawings. Each category consists of approximately 200 samples, indicating a balanced dataset. This balanced distribution is crucial for training machine learning models, as it helps prevent biases towards any one class and ensures that the model learns to accurately identify and differentiate between the two categories. A balanced dataset presents in Fig. 5 like this enables more reliable and generalizable results in predictive modeling, making it an ideal basis for conducting accurate analyses on children's drawing styles and their implications for developmental assessments.

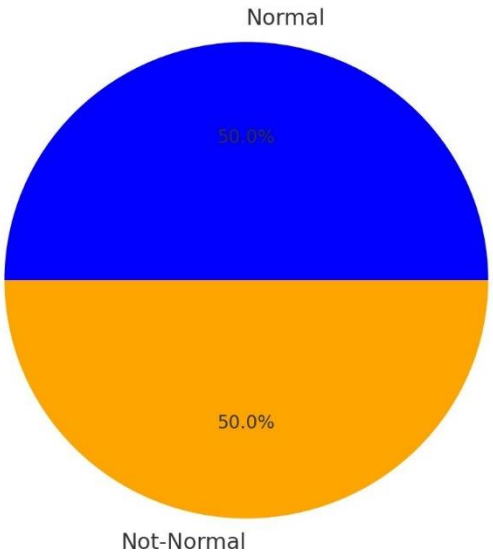

Fig. 5. "normal" and "not-normal" classes for children's hand drawings

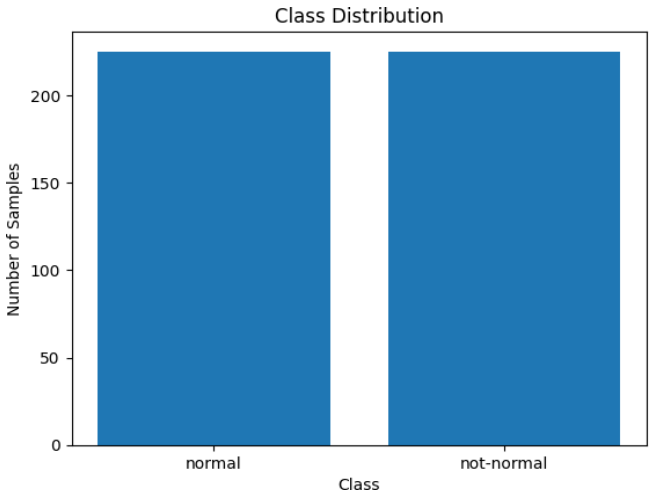

Fig. 6. Class distribution

C. Data Preprocessing

Preprocessing is a critical step in the machine learning pipeline, essential for preparing raw data into a clean dataset that enhances the performance and effectiveness of the models [40][41].

In our data preprocessing phase, we meticulously prepared our dataset for the training and testing of our deep learning models. The dataset is made up of hand-drawn images, which are separated into training and test or validation sections. The training set was made up of 80% of the images while the test/validation set was composed of the remaining 20%. Specifically, our split involved 360 images making up the training set; thus, ensuring representative diversity in both sets and 90 images set aside for test/validation. Image generation refers to the production of new images using algorithms. Two significant deep learning methods that are suited for image generation include Generative Adversarial Networks, or GANs [42], and Variational Autoencoders, or VAEs [43]. These networks are given a dataset to train on and learn the dataset's features and patterns, then create new images that preserve these features but with novel visual differences. Image generation in applications is used in increasing image quality or resolution enhancing and creating artistic images producing certain types of realistic human faces for use in animation, gaming, among other avenues and finally, creating synthetic datasets to use in ML architectures. These images are essential in fields where data images are high quality, and diversity is limited [44].

This was followed by more data processing using the ImageDataGenerator module from Keras that rescales the pixel values. Rescaling, referred to as normalization, achieves normalized pixel values from between 0 and 255 to between 0 and 1. The scale aims to simplify calculations by using lower numerical values during model training. It also published the image output to 240 by 240 pixels for standard image size in the neural network. Standardizing the image size allows the neural network to process all images equally and hence avoids biases or errors due to image size dimensions. For the training dataset, the input images were arranged in batches of 64 to optimize memory use while also enhancing the gradient estimation of the model during backpropagation. Lastly, the classification was set as binary for the two categories of normal and not-normal type images, focusing the model's training on distinguishing the two distinct categories.

Furthermore, in the same way, the test generator was set up to meet the likeness in image preprocessing and processing within the training and test steps. Therefore, this thoroughness in data preparation enhances model accuracy in addition to validating the model by allowing it to perform excellently independently of training data. This disciplined data preprocessing sets the stage for model training and evaluation.

D. Modeling

During the modeling stage of our research on classifying the hand drawings created by children, we carefully studied several deep learning frameworks to create wider ranging, more accurate, and dependable classifications. Our methodology consisted of two major components. The first one involved the development of unique models, while the second relied on up-to-date and recent pre-trained networks. In all instances, each model was adjusted to reflect the actual items and classification targets we had identified. Finally, we used an ensemble, a collection of distinct models, to improve the accuracy and dependability of our predictions.

- **CNN Architecture:** In this study, we employed a Convolutional Neural Network (CNN) specifically designed to differentiate between "normal" and "not-normal" children's drawings. Fig. 7 illustrates the CNN architecture. This CNN architecture was chosen based on its demonstrated effectiveness in image classification tasks, particularly for its ability to extract detailed features from complex image data. Research has shown that multiple convolutional layers followed by max-pooling layers effectively capture various levels of abstraction in images, making CNNs suitable for our application [45].

The architecture begins with a convolutional layer featuring 32 filters of size 3×3, utilizing the ReLU activation function to introduce nonlinearity, thereby enabling the model to learn complex patterns. The depth of the network increases with subsequent layers, which use 64 and then 128 filters. This progression is designed to incrementally capture more refined details in the images. After each convolutional layer, a max-pooling layer reduces dimensionality, helping to prevent overfitting by summarizing the features extracted in the convolutional layers.

To further mitigate overfitting, dropout layers with a rate of 0.5 are strategically placed after the dense layer of 512 neurons. This setup randomly deactivates a portion of the neurons during training, enhancing the model's ability to generalize to new, unseen data.

We conducted extensive hyperparameter tuning to optimize the model's performance. Parameters such as filter size, number of filters, and dropout rate were adjusted based on iterative testing and validation on a subset of the data. The final layer is a dense layer with a sigmoid activation function, tailored for binary classification tasks, providing the probability of a drawing being classified as "not-normal."

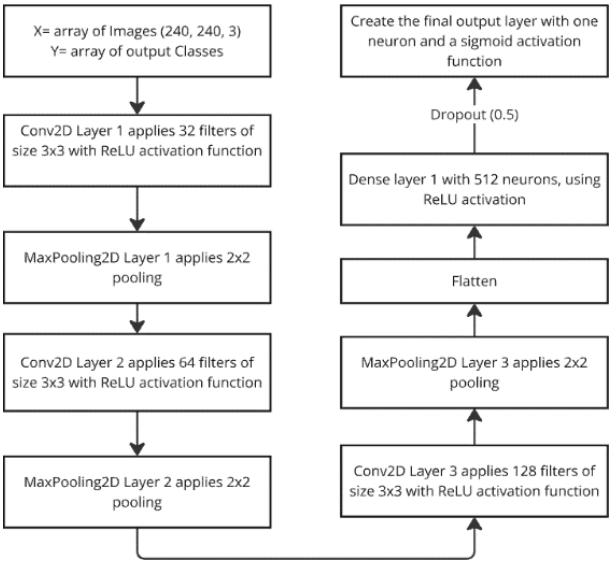

Fig. 7. The proposed CNN architecture

- **VGG16 Architecture:** The VGG16 network [46] architecture, depicted in Fig. 8, is comprised of 16 layers with trainable weights: 13 convolutional layers and 3 fully connected layers [47]. Each convolutional layeremploys filters with a small receptive field of 3×3—the minimum size required to capture directional nuances such as left/right and up/down. The stride for these filters is fixed at 1 pixel, and the input to each convolutional layer is padded by 1 pixel to ensure that the spatial resolution is preserved post-convolution.

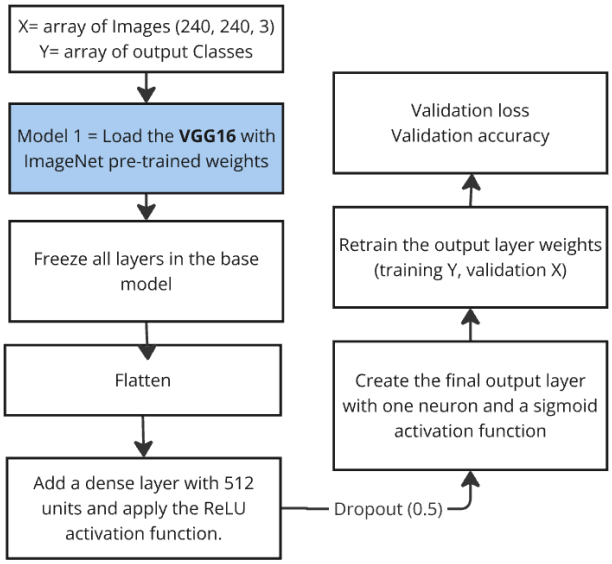

Fig. 8. The proposed VGG16 architecture

Following each convolutional operation, a ReLU activation function introduces non-linearities into the model, facilitating the learning of more complex patterns from the data. These convolutional layers are grouped into blocks of 2 or 3, each followed by a max-pooling layer with a 2×2 filter size and a stride of 2. This setup effectively reduces the dimensionality of the input volume, thus decreasing computation needs and helping control overfitting by summarizing the features extracted in the convolutional layers.

The architecture concludes with three fully connected layers positioned after multiple blocks of convolutional and max-pooling layers. Initially designed for large-scale multi-class classification, the first two layers contain 4096 nodes each, and the final layer traditionally has 1000 nodes configured for a 1000-way classification across various classes using a softmax activation. In the context of our study focusing on binary classification, this last softmax layer is modified to a simpler format with one node employing a sigmoid activation function to decide between two distinct classes: normal and not-normal. This tailored setup in VGG16 is specifically adjusted to enhance its applicability for binary tasks, as detailed in the accompanying figure, demonstrating how these adaptations are crucial for achieving precise outcomes in our specific classification objectives.

- **VGG19 Architecture:** The VGG19 [48] model, an extension of VGG16, is detailed in Fig. 9 and features 19 layers with trainable weights. This model includes 16 convolutional layers that are organized into more comprehensive blocks compared to VGG16, along with 3 fully connected layers. Similar to VGG16, it employs 3x3 convolutional filters but with more filtering layers added, increasing the depth of the network. It follows the same padding and stride strategies as VGG16 to maintain the spatial dimensions of the input through the convolutional layers. Each convolutional layer is followed by a ReLU activation function, and each block of convolutional layers is followed by a max-pooling layer with the same specifications as in VGG16, designed to reduce feature dimensions and to help in making the model invariant to small changes in the position of the feature in the input (see Fig. 9). Like VGG16 [49]–[51], VGG19's fully connected layers traditionally contain 4096, 4096, and 1000 neurons, respectively, ending in a softmax layer for classification across many classes. For tasks like ours involving binary classification, these are customized to end with a sigmoid function that outputs the probability of the input being in one of two classes (normal or not-normal).

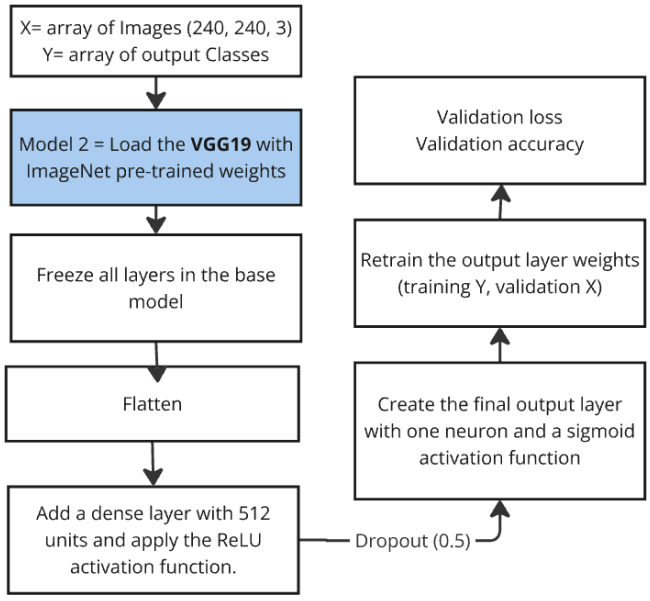

Fig. 9. The proposed VGG19 architecture

- **MobileNet Architecture:** MobileNetV2 [52]–[54] was chosen for its efficiency and performance in environments with limited computational resources, making it suitable for mobile applications. Similar to the adaptations made for VGG models as presents in Fig. 10, we included a non-trainable pretrained base followed by custom top layers. Instead of traditional pooling layers, MobileNetV2 utilizes a global average pooling layer to reduce spatial dimensions, which helps in maintaining the most important part of the feature maps. The subsequent layers include a dense layer with 256 neurons, a dropout layer to prevent overfitting, and a sigmoid output layer for binary classification.

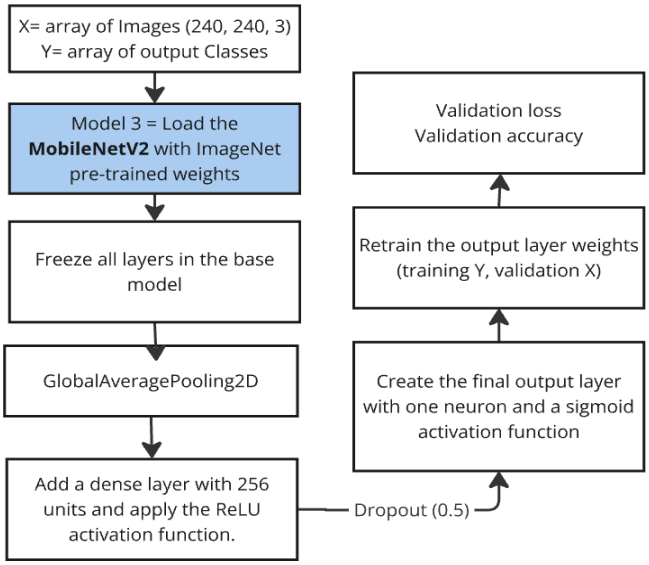

Fig. 10. The proposed MobileNet architecture

- **Ensemble Model - Hard Voting:** In response to the need for high accuracy and reliability in classifying children's drawings as either normal or not-normal, we have implemented a hard voting ensemble method, integrating the strengths of VGG16, VGG19, and MobileNet architectures. This ensemble approach leverages the unique advantages of each model: VGG16 and VGG19 for their deep, structured architectures known for high-performance in image classification tasks, and MobileNet for its efficiency in handling mobile-based applications with limited computational resources. The combination of these models aims to mitigate individual biases and errors, enhancing the robustness and objectivity of the final predictions.

Each model within our ensemble has been optimized using the Adam optimizer, renowned for its efficiency with sparse gradients and noisy data. We employ binary cross-entropy as our loss function, which is well-suited for binary classification problems, ensuring that our model effectively distinguishes between the two classes.

By harnessing the collective capabilities of these diverse architectures through a hard voting mechanism, where the final output is decided by the most common prediction across the models for each sample, our approach not only improves the accuracy but also the generalizability of the system across different sets of children's drawings. This ensemble method allows us to capitalize on the complementary strengths of each model, thereby providing a more reliable and powerful tool for psychological and developmental assessments based on children's drawings. Despite the complexity introduced by combining multiple models, the benefits in terms of enhanced predictive power and reduced risk of overfitting justify the approach, especially given the critical nature of the application in educational and developmental settings.

E. Evaluation Metrics

The evaluation phase of our deep learning models, which prioritize classifying children's drawings as either normal or not-normal. We use several key performance metrics that are essential to measure the models' efficacy. They ascertain the overall performance and how the model fares in terms of how it struggles with different kinds of classification errors. We use the confusion matrix, accuracy, precision, recall, and f1-score as our main metrics. These matrixes help us establish how well the models differentiate between the two categories and how they balance classification accuracy and the risk of misclassification.

- Confusion Matrix The confusion matrix is a fundamental evaluation tool that provides a detailed breakdown of the classification results for each class. It is typically structured as a table with four different outcomes:

– True Positives (TP): Correctly predicted positive observations.

– True Negatives (TN): Correctly predicted negative observations.

– False Positives (FP): Incorrectly predicted as positive (Type I error).

– False Negatives (FN): Incorrectly predicted as negative (Type II error).

The matrix itself helps in visualizing the performance of an algorithm.

1) Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observations to the total observations. It is great as a measure when the target classes are well balanced. The formula for accuracy is:

$$ACC = \frac{TN+TP}{TP+TN+FP+FN} \quad (1)$$

2) Precision

Precision [55]–[57] is the ratio of correctly predicted positive observations to the total predicted positives. It is a measure of the accuracy provided that a class label has been predicted, i.e., it quantifies the number of true positives over the sum of true positives and false positives.

$$PRE = \frac{TP}{FP+TP} \quad (2)$$

3) Recall

Recall [58]–[60] is the ratio of correctly predicted positive observations to all observations in actual class. It provides an indication of missed positive predictions.

$$REC = \frac{TP}{FP+TN} \quad (3)$$

4) Precision F1-Score

The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is especially useful when the classes are imbalanced. The F1 score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

$$F1 - S = 2 \times \frac{PRE \times REC}{PRE+REC} \quad (4)$$

To address the lack of detailed error analysis, we delve deeper into understanding the implications of false positives versus false negatives. In the context of developmental assessments, a false positive (incorrectly labeling a normal drawing as not-normal) could lead to unnecessary psychological evaluations, whereas a false negative (failing to identify a not-normal drawing) might result in a missed opportunity for early intervention. This nuanced understanding of error types helps in refining the model to reduce specific errors that could have significant implications for a child's developmental pathway. While our current focus is on binary classification, this approach may indeed oversimplify the complex spectrum of children's drawing behaviors. To enrich our model's assessment capabilities, future iterations could incorporate a more nuanced classification scheme that includes intermediate categories or utilizes additional metrics that capture broader psychological and developmental dimensions. This would allow for a more granular and accurate reflection of the diverse expressions evident in children's drawings, potentially leading to more tailored and informative developmental assessments.

We interpret these metrics to understand the balance between identifying drawings that may indicate psychological issues (recall) and minimizing the risk of false alarms (precision). High precision with balanced recall would mean our model effectively identifies drawings indicative of psychological concerns with minimal misclassification, guiding appropriate follow-up assessments.

Our study, involving children, adheres strictly to ethical guidelines. We ensure the confidentiality and anonymity of participant data, with appropriate consent from guardians. We also consider the psychological impact of our classifications, aiming to enhance, not replace, traditional psychological assessments with AI.

IV. EXPERIMENTAL RESULTS

All of the models have been trained via the Adam optimizer [58]–[61]. This was chosen as an alternative to the traditional stochastic gradient descent models, as the Adam optimizer is designed specifically for an efficient work with sparse gradients on noisy tasks, such as an image classification. The optimization method is an adaptive learning algorithm, which lead to both better and faster results than a non-adaptive approach. The epochs number for training was set to twenty.

The justification for this choice was the evaluation of early optimization training results, which allowed for the stable convergence of the system without overfitting, which is a frequent issue with deep learning AI.

The training was done utilizing the train generator function. The function permits real-time batch image augmentation while training to extend our model's potential to generalize. A test generator object was used together with the validation data parameter to supply a separate dataset to the model that it does not see during training. By setting the verbose parameter to 1, we could see the loss and accuracy for the training and validation data on the eve of each epoch. We could exercise our models in a strict environment and iterate on the outcomes cautiously to get the best achievable performance.

A. Results of Convolutional Neural Network

Training and validation metrics across 20 epochs of our CNN model provides a comprehensive insight into the learning and performance trends (see Fig. 11). The model seemed to struggle with the underlying data patterns, given the high training loss [65]– [68] of 1.1058 and low/training accuracy of 44.72%. However, a commonality of trends emerged throughout the epochs.

The graphs reveal a consistent decrease in both training and validation losses across the epochs, with the training loss dramatically dropping from 1.1058 to 0.0341 by the final epoch. Correspondingly, training accuracy soared to 98.89%, and validation accuracy stabilized at approximately 94.44%. These results demonstrate that the model not only learned effectively from the training data but also generalized well to the validation set.

Despite the convergence of training and validation losses and high accuracy rates, the potential for overfitting cannot be dismissed without more robust validation. Although the similar trajectories of training and validation metrics suggest a well-tuned model configuration, further analysis involving additional tests on unseen data sets or employing techniques such as cross-validation could provide stronger evidence of the model's ability to generalize. Moreover, exploring the impact of training duration and the number of epochs on model performance could offer deeper insights. Assessing whether extending beyond 20 epochs leads to diminishing returns or further performance enhancements would help optimize the training strategy, ensuring the model's reliability and robustness in categorizing children's drawings into 'normal' and 'not-normal' categories effectively.

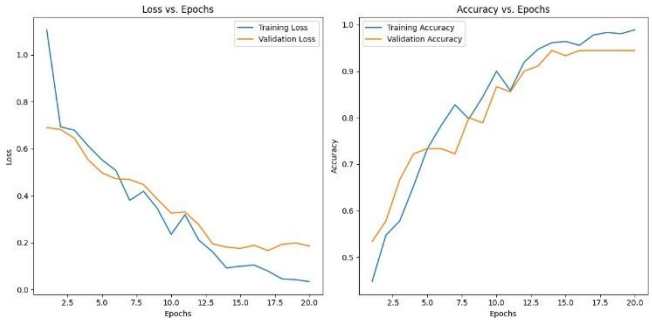

Fig. 11. Learning curves of CNN

The confusion matrix [69]–[73] for our Convolutional Neural Network (CNN), as shown in Fig. 12, provides a detailed look at the model's effectiveness in classifying children's drawings into 'normal' and 'not-normal' categories. The model achieves a high accuracy level, demonstrated by the presence of 44 true positives and 41 true negatives, effectively distinguishing between the two classifications. This accuracy is also highlighted by the minimal misclassifications observed, with only 1 false negative and 4 false positives reported. This indicates that the model is particularly adept at correctly identifying 'normal' drawings, an essential trait for applications where accurate identification of normal conditions is vital. However, the occurrence of 4 false positives indicates a slight tendency of the model to misclassify 'not-normal' drawings as 'normal', signaling an area for potential improvement. Despite this, the high rates of true positives and true negatives, combined with low numbers of false classifications, underline the model's robustness and reliability. This confirms its effectiveness in generalizing from the training data, making it a valuable asset for automated analysis in practical applications. The confusion matrix for our Convolutional Neural Network (CNN), as shown in Fig. 12, provides a detailed look at the model's effectiveness in classifying children's drawings into 'normal' and 'not-normal' categories. The model achieves a high accuracy level, demonstrated by the presence of 44 true positives and 41 true negatives, effectively distinguishing between the two classifications. This accuracy is also highlighted by the minimal misclassifications observed, with only 1 false negative and 4 false positives reported. This indicates that the model is particularly adept at correctly identifying 'normal' drawings, an essential trait for applications where accurate identification of normal conditions is vital. However, the occurrence of 4 false positives indicates a slight tendency of the model to misclassify 'not-normal' drawings as 'normal', signaling an area for potential improvement. Despite this, the high rates of true positives and true negatives, combined with low numbers of false classifications, underline the model's robustness and reliability. This confirms its effectiveness in generalizing from the training data, making it a valuable asset for automated analysis in practical applications.

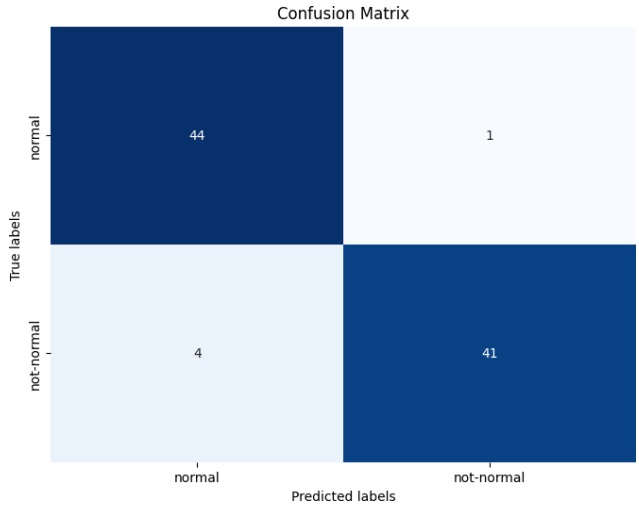

Fig. 12. Confusion matrix of CNN

Furthermore, the presence of only one false negative is indicative of the model's strong sensitivity, essential for ensuring that nearly all 'not-normal' conditions are detected. However, the impact of these errors needs to be considered more deeply. False positives, while less critical than false negatives in many health-related applications, could lead to undue anxiety or unnecessary testing in a psychological assessment context. Conversely, the rare false negatives, though minimal, are significant in that even a single missed detection of a 'not-normal' condition could mean a missed opportunity for early intervention.

The CNN model demonstrates excellent performance in classifying children's drawings into 'normal' and 'not-normal' categories, as evidenced by the high test accuracy of approximately 94.44% This high level of accuracy reflects the model's strong ability to generalize from the training data to unseen test data, a key indicator of a well-trained machine learning model.

B. Results of VGG16

As demonstrated in Fig. 13 of loss and accuracy plots during 20 epochs, the performance of the VGG16 model is characterized by impressive learning and generalization skills. At the very beginning, the model started with a high training loss of 3.2304; however, it was rapidly reduced, which proves its ability to quickly adapt to the features of the training dataset. By the second epoch, the training loss had already reached the level of 1.4845 visibility. The decrease in the validation loss is also seen at this stage, which means that the model was able to generalize sufficiently to predict unseen data. The trends on the accuracy plot are similar: the model's percentage on the training set started at 59.17% and quickly increased to 72.22% until the second epoch. By the 12th epoch, it reached 100%; this means that the model has truly understood the data and was able to perfectly predict the training set. The validation accuracy score has a similar pattern and after the ninth epoch stabilized at around 97.78%. Despite these positive indicators, the later epochs showed some fluctuations in validation loss, although the sustained high accuracy suggests that these did not compromise the model's ability to generalize. This performance, characterized by nearly flawless accuracy and minimal loss on unseen data, underscores the VGG16 model's efficacy in handling complex image classifications, such as those required for analyzing children's drawings.

However, in response to concerns about potential overfitting given the high accuracy and low loss, and the adequacy of the epoch number, further analysis would be beneficial. Exploring whether extending the training beyond 20 epochs would result in diminishing returns or continued improvements could offer deeper insights into the training dynamics and help validate the chosen epoch number. Implementing more robust measures, such as cross-validation or additional unseen datasets, could also ensure that the model not only fits the training data but also generalizes well to new, diverse data scenarios, thus enhancing the model's practical applicability in real-world settings.

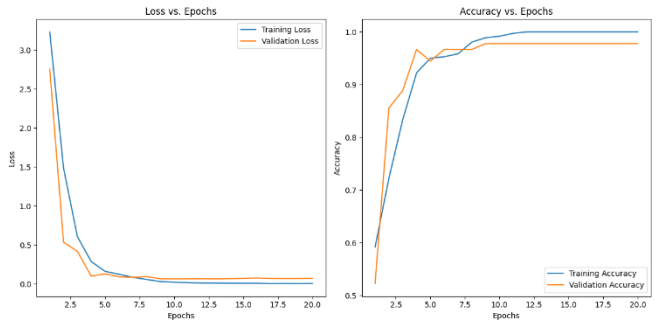

Fig. 13. Learning curves of VGG16

Another compelling visualization of the accuracy of the VGG16 model in classifying children's drawings into 'normal' and 'not-normal' shows the confusion matrix in Fig. 14. It is evident that the model was highly precise, with 44 true positives and 44 true negatives. In other words, the VGG16 model performed with high accuracy in identifying the drawings well. Such precision is important for any automated classification prediction system to be dependent. The good performance of the VGG16 model on both classes is even more important. The near-symmetrical distribution of errors within the confusion matrix underscores the model's unbiased nature in prediction. This lack of bias is crucial for maintaining the integrity of the classification system, ensuring that one category is not unduly prioritized at the expense of another. A deeper examination of these error types—false positives and false negatives—reveals their respective impacts on the practical application of the model. While false positives might lead to unnecessary follow-up assessments, false negatives could potentially overlook critical abnormal conditions that require intervention. Understanding the implications of these errors is fundamental to refining the model's accuracy and reliability in real-world applications, highlighting areas for potential enhancement to further optimize the system's performance.

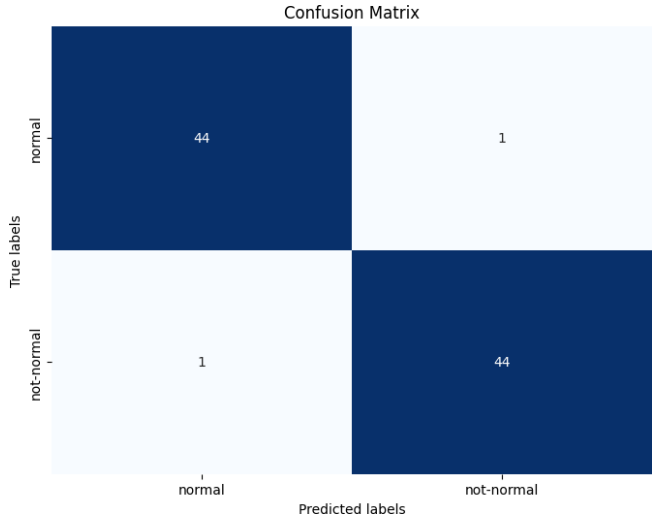

Fig. 14. Confusion matrix of VGG 16

The model is highly proficient in distinguishing between children's drawings which are 'normal' from the ones labeled 'not-normal'. This level of data convincingly proves the model's effectiveness in differentiating the two classes and makes it dependable in use cases for image classification where excellent accuracy is critical. A testing accuracy of almost 97.78% is also ranked high and, as such, shows that the model can generalize effectively from the learning set to the previously unseen testing data. An accuracy degree of the above ratio verifies that VGG16 can provide consistently accurate prediction for different data sets.

C. Results of VGG19

The VGG19 model performance over 20 epochs, as shown in Fig. 15, illustrates a successful training and validation progression, characterized by substantial improvement in both accuracy and loss reduction. Initially, the model started with a high training loss of 3.1686, which sharply decreased to 1.0944 by the second epoch and continued to decline steadily, settling at 0.0146 by the 20th epoch. This rapid decrease in loss indicates that the model efficiently learned the distinguishing features of the data early in the training process.

Similarly, the accuracy metrics exhibit a notable improvement from the outset. The training accuracy began at 57.22% and rapidly increased to over 94% by the ninth epoch, eventually achieving a perfect accuracy rate of 100% from the 15th epoch onward. The validation accuracy also displayed a consistent upward trend, beginning at 77.78% and rising to an impressive 97.78% by the final epoch, suggesting that the model was not only fitting well to the training data but was also highly effective at generalizing to new, unseen data.

Notably, the graphs also reveal a temporary increase in validation loss during the early epochs, peaking at 0.9080 at the second epoch. However, this issue was quickly resolved, and subsequent epochs saw a decrease and stabilization in validation loss, which closely aligned with the training loss from the tenth epoch onwards. This convergence of training and validation losses, along with consistently high accuracy levels, suggests that the model was well-optimized and balanced, displaying no overt signs of overfitting despite achieving high training accuracies.

To address potential concerns about overfitting and the appropriateness of the 20-epoch training duration, further analysis could be beneficial. Investigating whether extending training beyond 20 epochs leads to diminishing returns or further improvements, and validating the chosen epoch number through additional experiments, could provide a more comprehensive understanding of the model's training dynamics and its capacity to generalize effectively across varied and potentially novel datasets.

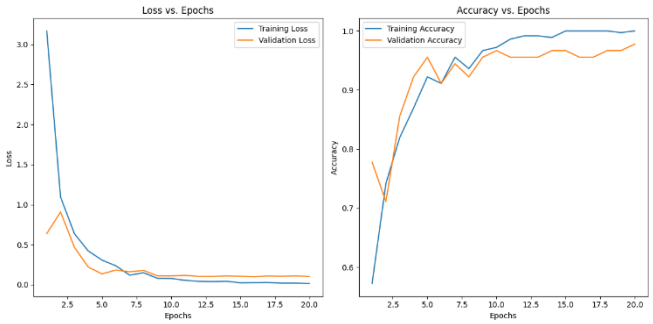

Fig. 15. Learning curves of VGG19

The confusion matrix in Fig. 16 of the VGG19 model gives an accurate picture of the model's capacity to classify children's drawings into normal and not-normal drawings. According to the matrix, the true positive rate of the normal and not-normal drawings was 44 for each. That is, the model achieved a high sensitivity and specificity, which is thrust upon near-perfect accuracy. Here are minimal errors assessing only one participant each for the false positive and false negative errors. It shows that the VGG19 model maintains a balanced sensitivity and specificity, achieving a recognition rate of the two classes with minimal bias to either category. The ultralow error rates confirm the negligible chances of misclassification, which is crucial in sensitive classifications. The automatic classification of normal and not-normal drawings in this research has shown that the VGG19 model is generally ideal in practical applications. This is considering its capability to deliver consistent accuracy in varied cases of normal and not-normal classifications.

The nuanced analysis of these error types, particularly the implications of the false positives and false negatives, is critical for understanding the practical impact of the model's performance. In contexts where the classification of children's drawings can influence further psychological assessment or educational interventions, minimizing errors is paramount. False positives might lead to unnecessary concern or additional testing, whereas false negatives could potentially overlook crucial developmental issues that require attention.

The performance dynamics of the VGG19 model, depict the strong capabilities of this model in classifying children's drawings as either 'normal' or 'not-normal.' The test accuracy rate of about 97.78% is a clear indicator of how true the models are in regards to their predictions within the set of data.

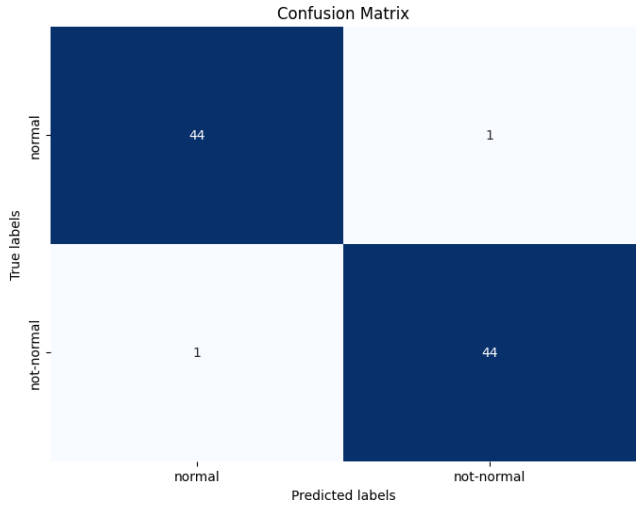

Fig. 16. Confusion matrix of VGG19

D. Results of MobileNet

The MobileNet model's performance over 20 epochs, as illustrated in Fig. 17, showcases a successful progression in terms of both training and validation metrics. The model started with a relatively high training loss of 0.5136, but this quickly declined, stabilizing at a low of 0.0320 by the final epoch. This rapid and substantial reduction in training loss indicates that the model effectively optimized its parameters to fit the training data well.

In parallel, the validation loss mirrored this positive trend, starting at 0.2742 and decreasing to 0.0716 by the 20th epoch. This consistency in the decrease of validation loss alongside the training loss indicates good generalization capabilities of the model, suggesting that it is not just fitting to the training data but also performing well on unseen data.

Accuracy trends further confirm the model's efficiency. Starting from an accuracy of 72.22%, the model quickly improved, reaching above 95% by the fifth epoch and peaking at 97.78% in the 15th epoch. This high level of accuracy, main-tained throughout the training process, showcases the model's robust capability to classify images accurately. The validation accuracy remained stable and high, predominantly hovering around 95-96%, and reaching up to 97.78%, underscoring the model's reliability and the effectiveness of MobileNet's architecture for the task at hand.

To address potential concerns about overfitting and the justification for the 20-epoch training regimen, further evaluations could be advantageous. Investigating whether extending the training duration beyond 20 epochs might result in diminishing returns or additional improvements could offer deeper insights into the model's training dynamics. Additionally, implementing more robust validation measures, such as cross-validation or testing on a broader and more diverse dataset, would help confirm the model's ability to generalize across different data scenarios and ensure its practical applicability in real-world settings.

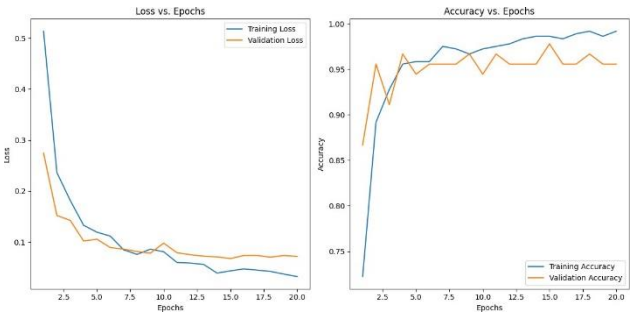

Fig. 17. Learning curves of MobileNet

The confusion matrix depicted in Fig. 18 for the MobileNet model demonstrates a high degree of accuracy in classifying children's drawings into 'normal' and 'not-normal' categories, with only a few misclassifications. The matrix shows that the model correctly identified 43 out of 45 'normal' drawings and 43 out of 45 'not-normal' drawings. This results in a substantial number of true positives and true negatives, underscoring the model's robustness in accurately classifying the two categories.

Despite this high accuracy, the model encountered some errors, specifically two instances each of false positives and false negatives. While these error counts are relatively low, they are crucial for understanding potential areas of improvement for the model. False positives—where 'not-normal' drawings are incorrectly classified as 'normal'—and false negatives—where 'normal' drawings are incorrectly

classified as 'not-normal'—can have differing implications depending on the application context.

In settings where accurate identification is critical, such as educational or psychological assessments, these errors could lead to inappropriate interventions or missed opportunities for support. A more nuanced analysis of these errors might involve exploring whether certain features of the drawings consistently lead to misclassification, or if variability within the 'normal' or 'not-normal' categories is not adequately captured by the current training dataset.

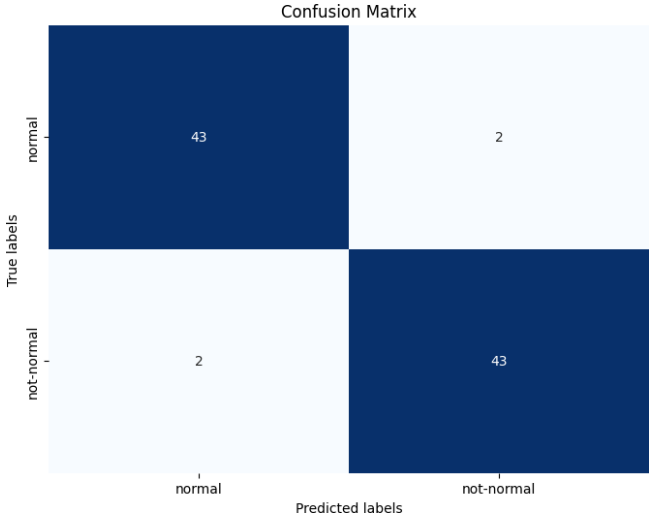

Fig. 18. Confusion matrix of MobileNet

E. Results of MobileNet

Our ensemble model strategically incorporates three of the top-performing models: VGG16, VGG19, and MobileNet, each selected for their distinct architectural features and learning capabilities that contribute to a robust ensemble method. VGG16 and VGG19, both from the VGG family, are renowned for their deep convolutional networks which facilitate the extraction of complex hierarchical image patterns. The additional layers in VGG19 allow for more detailed feature extraction compared to VGG16, offering slightly varied perspectives on the data processed. MobileNet, on the other hand, introduces a different approach by utilizing depthwise separable convo-lutions which optimize computational efficiency and reduce processing costs.

The rationale behind integrating these models into a single ensemble was to leverage their collective strengths to mitigate their individual limitations, thereby enhancing overall prediction accuracy and reliability. This method is particularly beneficial in complex classification scenarios such as ours, where distinguishing subtle differences between 'normal' and 'not-normal' categories is crucial. By harnessing the varied capabilities of these models, the ensemble approach aims to deliver a more consistent and unbiased performance across diverse input datasets, making it well-suited for practical applications like classifying children's drawings.

The effectiveness of this ensemble is demonstrated in the confusion matrix shown in Fig. 19, where it exhibits exemplary performance in classifying drawings as 'normal' and 'not-normal'. It successfully identified all 45 'normal' drawings with zero false negatives and accurately classified 44 out of 45 'not-normal' drawings, with just one false negative.

However, it is important to address potential drawbacks of using an ensemble method, such as increased computational complexity and the risk of overfitting. While ensemble models generally improve prediction accuracy, they require more computational resources and can be more complex to train and optimize. Moreover, the risk of overfitting can be heightened if not carefully managed, particularly when combining multiple high-performing models. Detailed analysis of the types of errors, such as the impact of false positives versus false negatives, is crucial for understanding the practical implications of these model errors. In practical applications, especially in sensitive fields like pediatric psychological assessment, the implications of each type of error must be carefully considered to ensure that the model's use aligns with clinical needs and ethical standards.

The ensemble hard voting model showcases its impressive performance. The model achieved an overall accuracy of approximately 98.89%, highlighting its efficacy in distinguishing between 'normal' and 'not-normal' categories in children's drawings. This high level of accuracy demonstrates the model's robustness and its strong potential to generalize to new or unseen data.

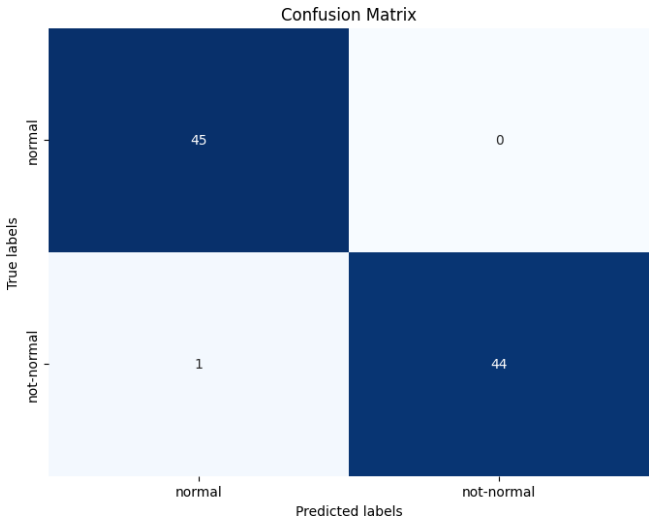

Fig. 19. Confusion matrix of ensemble hard voting

V. COMPARATIVE ANALYSIS

The comparative analysis of performance metrics for the CNN, VGG16, VGG19, MobileNet models, and their ensemble using a hard voting method, as seen in Table II, provides significant insights into the efficacy of ensemble learning in machine learning applications, particularly in complex and high-dimensional challenges like image classification. Each model exhibits commendable performance on its own. For example, both VGG16 and VGG19 demonstrate exceptional precision, recall, and F1-score values of 0.98, affirming their capabilities in handling image classification tasks with considerable complexity. MobileNet, tailored for mobile applications that require resource efficiency, consistently delivers a performance

metric of 0.96, highlighting its balance between accuracy and resource utilization. CNN shows robust performance, particularly with a high recall rate for the normal class, suggesting it effectively identifies true positive cases.

The ensemble approach, utilizing the hard voting method among CNN, VGG16, VGG19, and MobileNet, elevates these individual performances by combining their strengths, aiming to counteract their individual weaknesses. This strategic integration leads to an ensemble model that achieves nearperfect accuracy of 0.99 and an outstanding balance of recall and precision. The selection of these specific models for the ensemble was driven by their complementary strengths and distinct architectural designs, which, when combined, minimize the errors each model might make individually, thus enhancing generalization on new, unseen data and reducing the risks associated with overfitting.

This nuanced error analysis is crucial for understanding the practical impact of model errors. For instance, the minimal false negatives and false positives indicate a highly reliable system, yet even a single error in a real-world application can have significant implications, especially in sensitive settings such as psychological assessments of children. Future work will delve deeper into how the characteristics of the Kids' Hand Movement Dataset (KHMD) might influence model performance and generalizability to ensure robustness across broader applications. Furthermore, the real-world applicability and interpretability of the ensemble model will be explored to maximize its utility in practical settings, ensuring that it not only performs well under experimental conditions but also in real-world deployments where nuanced interpretation and adaptability are essential.

The comparative analysis of our ensemble model with state-of-the-art models across various studies reveals several significant insights, as summarized in Table III. The research conducted by Pysal et al. utilized a FLSTM model on a private drawing dataset, achieving an accuracy and F1-score of 88.6%. This performance is commendable, particularly in the context of children's drawing strategies and seriation objects, indicating a robust model for sequencing and stroke order analysis. Shi et al.'s work with an ASD classifier on the Paintings of Autism Spectrum Disorder (PASD) dataset reported an accuracy of 85.45%, showcasing the model's effectiveness in distinguishing paintings by ASD individuals from typically developed children. However, the absence of an F1-score limits a comprehensive understanding of the model's balance between precision and recall.

Kamran et al. achieved an impressive accuracy of 99.22% using a finetuned AlexNet on a combined dataset comprising HandPD, NewHandPD, and Parkinson's Drawing datasets. This highlights the potential of deep learning models in early-stage Parkinson's disease detection through handwriting analysis, although the F1-score was not reported, which is crucial for understanding the model's robustness in classification tasks. Elngar et al. combined ANN and CNN models to evaluate personality traits on the Personality Detection.

Dataset (PDD), achieving an accuracy of 70.73%. While this indicates a significant improvement over traditional methods, the model's overall performance is lower compared to other state-of-the-art approaches.

Ghosh et al. employed a graphology-based analysis method on a newly collected dataset, achieving an F1-score of 86.70%. This demonstrates the model's capability in analyzing human behavioral characteristics through handwriting, albeit without an accuracy metric, which would have provided a more rounded evaluation of the model's performance. Mekhaznia et al. used an ANN on the TxPI-u dataset, achieving an accuracy of 71.90%, indicating reasonable performance in personality trait identification among undergraduates, though still not on par with the highest-performing models.

TABLE II. COMPARATIVE PERFORMANCE OF MODELS

Model	Precision		Recall		F1-Score		Accuracy
	Normal	Not-Normal	Normal	Not-Normal	Normal	Not-Normal	
CNN	0.92	0.98	0.98	0.91	0.95	0.94	0.94
VGG16	0.98	0.98	0.98	0.98	0.98	0.98	0.98
VGG19	0.98	0.98	0.98	0.98	0.98	0.98	0.98
MobileNet	0.96	0.96	0.96	0.96	0.96	0.96	0.96
Ensemble Hard Voting	0.98	1.00	1.00	0.98	0.99	0.99	0.99

TABLE III. COMPARATIVE STUDY BETWEEN OUR WORK AND PREVIOUS WORKS

Author	Model	Dataset	Accuracy	F1-score
Pysal et al. [27]	FLSTM	private drawing dataset	88.6	88.6%
Shi et al. [29]	ASD Classifier	Paintings of Autism Spectrum Disorder (PASD)	85.45%	-
Kamran et al. [30]	AlexNet finetuned	combined HandPD, NewHandPD and Parkinson's Drawing datasets	99.22%	-
Elngar et al. [33]	ANN + CNN	Personality Detection Dataset (PDD)	70.73%	-
Ghosh et al. [34]	Graphology-based Analysis	new collected dataset	86.70	-%
Mekhaznia et al. [35]	ANN	baptised Text for Personality Identification of Undergraduates, baptised TxPI-u	71.90%	-
Ours	Ensemble Hard voting	Kids' Hand Movement Dataset (KHMD)	99%	99%

Our method, employing an ensemble hard voting approach on the Kids' Hand Movement Dataset (KHMD), achieved both an accuracy and F1-score of 99%. This remarkable performance underscores the effectiveness of integrating multiple models, such as custom CNN, VGG16, VGG19, and MobileNet, to leverage their collective strengths while mitigating individual weaknesses. The high accuracy and F1-score indicate that our ensemble model not only excels in precision but also maintains a balanced recall, ensuring reliable classification of children's hand drawing movements. This robust performance suggests that our methodology is wellsuited for applications in psychological and developmental assessments, providing a significant advancement over existing models in terms of accuracy and reliability. The results demonstrate that our ensemble approach can effectively address the biases and limitations inherent in individual models, leading to superior overall performance in classifying and understanding children's drawing behaviors.

VI. CONCLUSION

To conclude, our study successfully demonstrates that deep learning models such as CNN, VGG16, VGG19, and MobileNet can effectively classify children's drawings into normal and not-normal categories. Notably, the ensemble model utilizing a hard voting mechanism outperformed the individual models, achieving an impressive accuracy of 99%. This superior performance illustrates the benefit of combining models with unique strengths and weaknesses to enhance overall predictive power and reliability.

For future work, we plan to expand our dataset to include a wider range of drawing styles and age groups to further validate and improve our model's robustness. Additionally, implementing more sophisticated ensemble techniques such as stacked generalization or blending may provide deeper insights and further performance enhancements. Exploring the application of these models in real-time drawing classification systems could also be beneficial for educational and psychological assessments, providing immediate feedback and support to children based on their drawings. This approach holds significant potential to refine educational tools and aid in early psychological intervention, impacting both education and child development profoundly.

REFERENCES

[1] E. A. Rider, E. Ansari, P. H. Varrin, and J. Sparrow, "Mental health and wellbeing of children and adolescents during the covid-19 pandemic," *bmj*, vol. 374, 2021.

[2] C. S. M. Ng and S. S. L. Ng, "Impact of the COVID-19 pandemic on children's mental health: A systematic review," *Front. Psychiatry*, vol. 13, p. 975936, Oct. 2022.

[3] I. Braito, T. Rudd, D. Buyuktaskin, M. Ahmed, C. Glancy, and A. Mulligan, "Review: systematic review of effectiveness of art psychotherapy in children with mental health disorders," *Ir. J. Med. Sci.*, vol. 191, no. 3, pp. 1369–1383, Jun. 2022.

[4] Z. Amod, R. Gericke, and K. Bain, "Projective assessment using the draw-a-person test and kinetic family drawing in south africa," *Psychological*, p. 375, 2013.

[5] W. A. Bainbridge, "A tutorial on capturing mental representations through drawing and crowd-sourced scoring," *Behav. Res. Methods*, vol. 54, no. 2, pp. 663–675, Apr. 2022.

[6] O. Altun and O. Nooruldeen, "Sketrack: Stroke-based recognition of online hand-drawn sketches of arrow-connected diagrams and digital logic circuit diagrams," *Scientific Programming*, vol. 2019, no. 1, p. 6501264, 2019.

[7] R. R. Rachala and M. R. Panicker, "Hand-drawn electrical circuit recognition using object detection and node recognition," *SN Computer Science*, vol. 3, no. 3, p. 244, 2022.

[8] K. Wrobel, R. Doroz, P. Porwik, T. Orczyk, A. B. Cavalcante, and M. Grajzer, "Features of hand-drawn spirals for recognition of parkinson's disease," in *Asian Conference on Intelligent Information and Database Systems*, pp. 458–469, 2022.

[9] S. Roy, A. Bhattacharya, N. Sarkar, S. Malakar, and R. Sarkar, "Offline hand-drawn circuit component recognition using texture and shapebased features," *Multimedia Tools and Applications*, vol. 79, pp. 353–373, 2020.

[10] M. Gupta, P. Mehndiratta, and A. Bhardwaj, "Object recognition in hand drawn images using machine ensembling techniques and smote sampling," in *Information, Communication and Computing Technology: 4th International Conference, ICICCT 2019*, pp. 228–239, 2019.

[11] M. Gupta and P. Mehndiratta, "Analysis and recognition of handdrawn images with effective data handling," in *Big Data Analytics: 7th International Conference, BDA 2019*, pp. 389–407, 2019.

[12] W. Adorno, A. Yi, M. Durieux, and D. Brown, "Hand-drawn symbol recognition of surgical flowsheet graphs with deep image segmentation," in *2020 IEEE 20th international conference on bioinformatics and bioengineering (BIBE)*, pp. 295–302, 2020.

[13] J. Adhikari, M. Aththanayake, C. Kularathna, A. Wijayasiri, and A. Munasinghe, "Deep learning based hand-drawn molecular structure recognition and 3d visualisation using augmented reality," in *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 31–38, 2022.

[14] M. S. Thangakrishnan and K. Ramar, "Retracted article: Automated hand-drawn sketches retrieval and recognition using regularized particle swarm optimization based deep convolutional neural network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 6, pp. 6407–6419, 2021.

[15] S. Ali, N. Aslam, D. Kim, A. Abbas, S. Tufail, and B. Azhar, "Context awareness based sketch-deepnet architecture for hand-drawn sketches classification and recognition in aiot," *PeerJ Computer Science*, vol. 9, p. e1186, 2023.

[16] Z. Li, J. Yang, Y. Wang, M. Cai, X. Liu, and K. Lu, "Early diagnosis of parkinson's disease using continuous convolution network: Handwriting recognition based on off-line hand drawing without template," *Journal of biomedical informatics*, vol. 130, p. 104085, 2022.

[17] X. Hou, X. Rong, and X. Yu, "Light-srnet: a lightweight dual-attention feature fusion network for hand-drawn sketch recognition," *Journal of Electronic Imaging*, vol. 32, no. 1, pp. 013 005–013 005, 2023.

[18] A. Keerthi Priya, N. Gaganashree, K. Hemalatha, J. S. Chembeti, and T. Kavitha, "Ai-based online hand drawn engineering symbol classification and recognition," in *Innovations in Electronics and Communication Engineering: Proceedings of the 9th ICIECE 2021*, pp. 195–204, 2022.

[19] J. Singh, K. Upreti, A. K. Gupta, N. Dave, A. Surana, and D. Mishra, "Deep learning approach for hand drawn emoji identification," in *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)*, pp. 1–6, 2022.

[20] S. Dey, A. Dutta, J. Llados´s, A. Forne´s, and U. Pal, "Shallow neural network model for hand-drawn symbol recognition in multi-writer scenario," in *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, vol. 2, pp. 31–32, 2017.

[21] H. Cecotti, C. Boumedine, and M. Callaghan, "Hand-drawn symbol recognition in immersive virtual reality using deep extreme learning machines," in *Recent Trends in Image Processing and Pattern Recognition: First International Conference, RTIP2R 2016*, pp. 80–92, 2017.

[22] S. Hayat, K. She, Y. Yu, and M. Mateen, "Deep cnn-based features for hand-drawn sketch recognition via transfer learning approach," *Editorial Preface From the Desk of Managing Editor*, vol. 10, no. 9, 2019.

[23] L. Akter *et al.*, "Early identification of parkinson's disease from hand-drawn images using histogram of oriented gradients and machine

learning techniques," in *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, pp. 1–6, 2020.

[24] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] A. G. Howard *et al*., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[26] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, pp. 241–258, 2020.

[27] D. Pysal, S. J. Abdulkadir, S. R. Mohd Shukri, and H. Alhussian, "Classification of children's drawing strategies on touch-screen of seriation objects using a novel deep learning hybrid model," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 115–129, Feb. 2021.

[28] M. F. Ahmadsaraei, A. Bastanfard, and A. Amini, "Child psychological drawing pattern detection on OBGET dataset, a case study on accuracy based on MYOLO v5 and MResNet 50," *Multimed. Tools Appl.*, vol. 83, no. 13, pp. 283–313, Apr. 2024.

[29] F. Shi, W. Sun, H. Duan, X. Liu, M. Hu, W. Wang, and G. Zhai, "Drawing reveals hallmarks of children with autism," *Displays*, vol. 67, p. 102000, Apr. 2021.

[30] I. Kamran, S. Naz, I. Razzak, and M. Imran, "Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease," *Future Gener. Comput. Syst.*, vol. 117, pp. 234–244, Apr. 2021.

[31] J. Zhang, Y. Lee, T.-M. Chung, and H. Park, "Development of a Handwriting Drawings Assessment System for Early Parkinson's Disease Identification with Deep Learning Methods," in *Future Data and Security Engineering. Big Data, Security and Privacy*, pp. 484–499, 2023.

[32] K. Nadeem, M. Ahmad, and M. Asif Habib, "Emotional States Detection Model from Handwriting by using Machine Learning," *2022 International Conference on Frontiers of Information Technology (FIT)*, pp. 284-289, 2022, doi: 10.1109/FIT57066.2022.00059.

[33] A. A. Elngar, N. Jain, D. Sharma, H. Negi, A. Trehan, and A. Srivastava, "A Deep Learning Based Analysis of the Big Five Personality Traits from Handwriting Samples Using Image Processing," *Journal of Information Technology Management*, vol. 12, pp. 3–35, Dec. 2020.

[34] S. Ghosh, P. Shivakumara, P. Roy, U. Pal, and T. Lu, "Graphology based handwritten character analysis for human behaviour identification," *CAAI Trans. Intell. Technol.*, vol. 5, no. 1, pp. 55–65, Mar. 2020.

[35] T. Mekhaznia, C. Djeddi, and S. Sarkar, "Personality Traits Identification Through Handwriting Analysis," in *Pattern Recognition and Artificial Intelligence*, pp. 155–169, 2021.

[36] A. Saraswal and U. R. Saxena, "Personality Trait Prediction Using Handwriting Recognition with KNN," *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp. 551-555, 2022, doi: 10.1109/CISES54857.2022.9844344.

[37] Y. Hamdi *et al*., "Deep learned BLSTM for online handwriting modeling simulating the Beta-Elliptic approach," *Engineering Science and Technology, an International Journal*, vol. 35, p. 101215, 2022.

[38] Y. Hamdi, H. Boubaker, B. Rabhi, W. Ouarda, and A. Alimi, "Hybrid architecture based on rnn-svm for multilingual handwriting recognition using beta-elliptic and cnn models," *Authorea Preprints*, 2023.

[39] M. C. Data, M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Exploratory data analysis," *Secondary analysis of electronic health records*, pp. 185–203, 2016.

[40] C. Li, "Preprocessing methods and pipelines of data mining: An overview," *arXiv preprint arXiv:1906.08510*, 2019.

[41] V. C¸ etin and O. Yildiz, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale U¨ niversitesi Mu¨hendislik Bilimleri Dergisi*, vol. 28, no. 2, pp. 299–312, 2022.

[42] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[43] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. AlamedaPineda, "Dynamical variational autoencoders: A comprehensive review," *arXiv preprint arXiv:2008.12595*, 2020.

[44] K. Maharana, S. Mondal, and B. Nemade, "A review: Data preprocessing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022.

[45] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[46] G. M. Devi and V. Neelambary, "Computer-aided diagnosis of white blood cell leukemia using vgg16 convolution neural network," in *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 1064–1068, 2022.

[47] R. Kaur, R. Kumar, and M. Gupta, "Review on transfer learning for convolutional neural network," in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 922–926, 2021.

[48] B. S. Kolla, B. R. Reddy, S. V. Sahithi, and L. P. Madala, "Comparative analysis of vgg19, resnet50, and googlenet inception models for bci," *Researchsquare*, 2023.

[49] Z.-P. Jiang, Y.-Y. Liu, Z.-E. Shao, and K.-W. Huang, "An improved vgg16 model for pneumonia image classification," *Applied Sciences*, vol. 11, no. 23, p. 11185, 2021.

[50] H. Qassim, D. Feinzimer, and A. Verma, "Residual squeeze vgg16," *arXiv preprint arXiv:1705.03004*, 2017.

[51] S. Mascarenhas and M. Agarwal, "A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification," in *2021 International conference on disruptive technologies for multidisciplinary research and applications (CENTCON)*, vol. 1, pp. 96–99, 2021.

[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[53] K. Dong, C. Zhou, Y. Ruan, and Y. Li, "Mobilenetv2 model for image classification," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pp. 476– 480, 2020.

[54] Q. Xiang, X. Wang, R. Li, G. Zhang, J. Lai, and Q. Hu, "Fruit image classification based on mobilenetv2 with transfer learning technique," in *Proceedings of the 3rd international conference on computer science and application engineering*, pp. 1–7, 2019.

[55] B. J. Erickson and F. Kitamura, "Magician's corner: 9. Performance metrics for machine learning models," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e200126, 2021.

[56] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Computer Science On-line Conference*, pp. 15–25, 2023.

[57] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

[58] T. Kynka¨a¨nniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *Advances in neural information processing systems*, vol. 32, 2019.

[59] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, "The area under the precision-recall curve as a performance metric for rare binary events," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 565–577, 2019.

[60] P. Mu¨ller, M. Brummel, and A. Braun, "Spatial recall index for machine learning algorithms," in *London Imaging Meeting*, vol. 2, pp. 58–62, 2021.

[61] R. Poojary and A. Pai, "Comparative study of model optimization techniques in fine-tuned cnn models," in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 1–4, 2019.

[62] E. M. Dogo, O. Afolabi, N. Nwulu, B. Twala, and C. Aigbavboa, "A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks," in *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*, pp. 92–99, 2018.

[63] D. O. Melinte and L. Vladareanu, "Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified adam optimizer," *Sensors*, vol. 20, no. 8, p. 2393, 2020.

[64] K. K. Kumar *et al*., "An efficient image classification of malaria parasite using convolutional neural network and adam optimizer," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 2, pp. 3376–3384, 2021.

[65] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, pp. 1–26, 2020.

[66] F. Nie, Z. Hu, and X. Li, "An investigation for loss functions widely used in machine learning," *Communications in Information and Systems*, vol. 18, no. 1, pp. 37–52, 2018.

[67] A. Jung. *Machine learning: the basics*. Springer Nature, 2022.

[68] Y. Zhang, J. Wen, G. Yang, Z. He, and J. Wang, "Path loss prediction based on machine learning: Principle, method, and data expansion," *Applied Sciences*, vol. 9, no. 9, p. 1908, 2019.

[69] J. T. Townsend, "Theoretical analysis of an alphabetic confusion matrix," *Perception & Psychophysics*, vol. 9, pp. 40–50, 1971.

[70] O. Caelen, "A bayesian interpretation of the confusion matrix," *Annals of Mathematics and Artificial Intelligence*, vol. 81, no. 3, pp. 429–450, 2017.

[71] N. D. Marom, L. Rokach, and A. Shmilovici, "Using the confusion matrix for improving ensemble classifiers," in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pp. 555–559, 2010.

[72] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection." *Maics*, vol. 710, no. 1, pp. 120–127, 2011.

[73] B. P. Salmon, W. Kleynhans, C. P. Schwegmann, and J. C. Olivier, "Proper comparison among methods using a confusion matrix," in *2015 IEEE International geoscience and remote sensing symposium (IGARSS)*, pp. 3057–3060, 2015.