

A DEEP LEARNING MODEL WITH A NEW LOSS FUNCTION FOR AGE ESTIMATION

ARMAN I. MOHAMMED, SARBAST H. ALI, OMER MOHAMMED SALIH HASSAN*
and SARDAR OMAR SALIH**

*Dept. of Information Technology, Duhok Polytechnic University, Kurdistan Region-Iraq

**Dept. of Web technology, Duhok Polytechnic University, Kurdistan Region-Iraq

(Received: September 19, 2023; Accepted for Publication: October 24, 2023)

ABSTRACT

Age estimation is a global challenge in the area of computer vision, as it depends on the facial features of the person. Recently, it has become an important approach for facial recognition problems and many other real-world applications. Accurate age estimation has the potential to improve decision-making processes in various industries and applications. It has been shown that the approach of Convolutional Neural Network (CNN) performs well for age estimation and promising results have been obtained by many researchers. However, the efficiency of CNN, to a great extent, is determined by the strength of the loss function. In this paper, a groundbreaking contribution is introduced, presenting a loss function that effectively calculates the disparity between the real and predicted age labels. Which is driven from Golden ratio and Mean Squared Error (MSE) functions. The proposed loss function is denoted by Golden Mean Squared Error (GMSE). A predesigned CNN is trained with UTKFace and FG-Net age datasets. According to the results, GMSE proved to operate better than preexisted loss functions. The MSE loss at epoch 25 was 51.34 and the GMSE loss at the same epoch was 3.15. At final round of training, the MSE loss was 6.56 and the GMSE loss was 1.58. The Mean Absolute Error (MAE) loss function was also used, but it couldn't lower the loss below 2 in the last epoch. Furthermore, the GMSE accuracy outperformed both MSE and MAE in the testing phase for both the UTKFace and FG-NET datasets. The GMSE loss function achieved better results than the MSE and MAE loss functions, indicating that it can save time and computations during the training process and provide better results at production phase.

KEYWORDS: Age Estimation, Convolutional Neural Networks, Deep Learning, Loss Function, Mean Squared Error.

1. INTRODUCTION

Age estimation has several potential benefits in the field of artificial intelligence. For instance, it can be used in facial recognition technology for security and law enforcement purposes, such as identifying suspects or missing persons. Additionally, it can be used in marketing and advertising to determine the age range of target audiences and tailor advertisements accordingly. Age estimation can also be used in healthcare for assessing patient age and predicting the risk of age-related diseases. Age estimation is the process of computing age from some information extracted

from human face. Developing a system for a precise age estimation is a challenging task and depends on some factors such as face expression, face features, posing, age features, lifestyle, light condition, etc., (Dhimar & Mistree, 2016; Tahir, 2012; Gupta & Nain 2023). However, the most difficult issue of age estimation is the extraction of age features from input facial images (Jun *et al.*, 2023). Also, the extra facial attachments such as glasses, beard, etc., for males and cosmetology and makeup for females represent obstacles for age estimation. (Figure 1) Shows the face of female with and without makeup. Generally speaking, there are two approaches for age estimation. The first approach includes the

prediction of age in an exact way by predicting the biological age (chronological) of a person from his/her facial image. The second approach includes the prediction of age in a form of age group, which is a less challenging task compared to the first approach, (Shen *et al.*, 2018; Han *et al.*, 2013). The first approach is categorized as a

regression problem while the second approach is categorized as a classification problem.

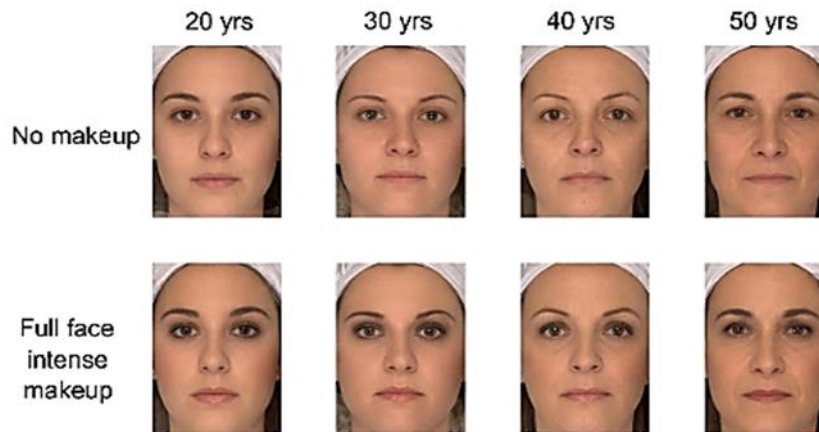


Fig. 1. The make-up effect on age, signs of aging can be visible with makeup when a person reaches the age of 50.

The estimated age is affected by many factors that make it far from the actual age. These factors are eyes, eye-brows, chin, cheek, nose, ears, face dimensions, rotation factors, gender, face orientation and dataset (Dhimar & Mistree, 2016;

Al-azzawi, 2021; Toshev & Szegedy, 2014). Some of these factors expand their areas as a result of the aging process, this growth is known as craniofacial growth which can be seen in (Figure 2).

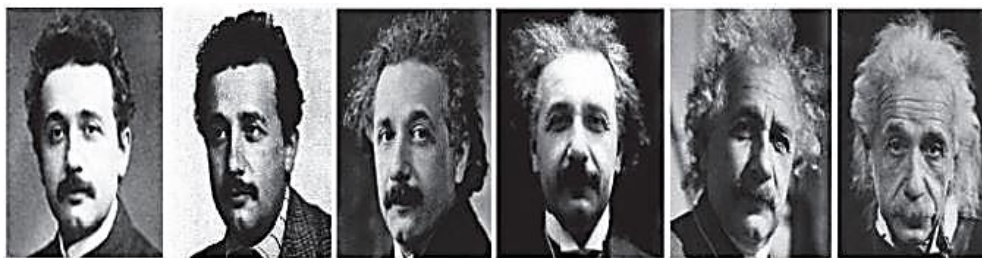


Fig. 2. Albert Einstein in a different era, where the craniofacial growth can be seen from nose, ears and also chin (Dhimar & Mistree, 2016).

Recently, the gap between estimated age and actual age has decreased to approximately 3 years using methods like biologically inspired features and support vector machines as in (Guo *et al.*, 2009). Thereafter, the Convolutional Neural Network (CNN) and support vector machine gives accuracy up to 96% for the 5 years age range (Dhimar & Mistree, 2016).

In this paper, the first approach is adopted, i.e., age estimation by regression method. Despite it is a more challenging approach, it provides better accuracy. Additionally, a loss function is developed to determine the loss of the predicted

age label, which is necessary to train the deep learning model effectively. Accurately calculating the loss is a crucial factor that directly affects the results of any deep learning model. Using an unsuitable loss function for a model can result in failure to converge and failure to minimize loss from the beginning. While several regression loss functions were intended for machine learning tasks, they can also function to a certain extent with conventional neural networks. However, specific tasks may still present some challenges. When training a convolutional regression network to estimate age

using classical regression loss functions like Mean Squared Error (MSE) or Mean Absolute Error, a substantial amount of loss will be generated during the initial stages of the training process (Lu *et al.*, 2018). This will result in high loss values while attempting to map between the actual age and predicted age, requiring extensive computations to train the models until the loss is minimized. Slow model processing at test time also will be a great issue. To overcome these limitations, and obtain the state-of-the-art model for the age estimation using convolutional neural networks the Golden Mean Squared Error (GMSE) loss function will be proposed in this paper and used along with the Soft Stagewise Regression Network (SSR NET) (Yang *et al.*, 2018).

2. LITERATURE REVIEW

In this paper the weight is given to the loss functions and their effect on deep learning models, here some of the most recent regression loss functions for deep learning and CNN structures for age estimation will be reviewed. Yang *et al.* proposed a loss function for salient object detection known as Progressive Self-Guided (PSG) loss function that mimics the morphological closing process on the model moderate predictions for making progressive and subsidiary training supervisions epoch-wisely (Yang *et al.*, 2021). Hu *et al.* designed a loss function for the age difference learning system to calculate age for images without labels, the proposed loss function was a combination of three losses including entropy loss, cross-entropy loss and K-L divergence distance loss (Hu *et al.*, 2017). Kendall and Cipolla proposed the geometric loss functions for camera pose regression with deep learning, the geometric loss was able to learn the weighting among translation and rotation automatically, using an estimate of the homoscedastic task uncertainty (Kendall & Cipolla, 2017). Later, in 2018, Xiankai and colleagues developed a shrinkage loss function to evaluate loss in convolutional regression tasks for object tracking. The shrinkage loss function computes the absolute difference from the squared loss (Lu *et al.*, 2018). Pan *et al.* proposed the mean-variance loss for age estimation task through distribution learning, by developing the joint loss function that contained softmax loss, variance loss and mean loss, their method yielded favorable results by reducing the Mean Absolute Error (MAE) in years, surpassing the

performance of other loss functions (Pan *et al.*, 2018). Zhang *et al.* during their study they proposed a Compact Cascade Context-based Age Estimation model (C3AE) a compact plain model by training facial images from three different resolutions low, medium and high (Zhang *et al.*, 2019). Liu *et al.* combined both tasks classification and regression, to obtain the outstanding model for the age estimation task (Liu *et al.*, 2020). Dornaika *et al.* studied the influence of loss functions in regression based CNNs for age estimation missions (Dornaika *et al.*, 2020). Bui *et al.* holds the view that loss functions play a significant role in the training process, and as a result, they opted to utilize contrastive and triplet loss functions to extract finer details from images (Bui *et al.*, 2018).

In addition, much work is done in developing the CNN neural network structure to reduce the MAE loss which indicates the difference between ground truth label and the predicted label. As well as increasing the performance of the estimation model. Berg *et al.* trained a deep multi-output convolutional neural network to classify training samples in multiple overlapping bins simultaneously by using randomized bins and obtained (4.55 MAE) which is a benchmark for age estimation using UTKFace dataset (Berg *et al.*, 2021). Yang *et al.* proposed a soft stagewise and compact regression network named SSR-Net and reduced the MAE loss for age estimation task to 3.16 using Morph2 dataset which was introduced in (Ricanek & Tesafaye, 2006) by developing a compact CNN for regression problems (Yang *et al.*, 2018). Agbo-Ajala and Viriri developed a CNN structure for age and gender classification that consists of six layers for feature extraction and classification which are four convolutional layers and two fully-connected layers, the network was trained over OIU- Adience dataset (Eidinger *et al.*, 2014), the exact accuracy obtained was 83.1 and One-off age was 93.8 (Agbo-Ajala & Viriri, 2020). Taheri and Toygar introduced a new type of deep neural network called Directed Acyclic Graph Convolutional Neural Networks (DAG-CNNs) for age estimation. This approach takes advantage of features from various layers of a CNN, using multiple stages to create more accurate predictions (Taheri & Toygar, 2019). Zhang *et al.* proposed a new Residual networks of Residual networks (RoR) CNN architecture (Zhang *et al.*, 2017), for high resolution age group and gender facial images classification and obtained 93.24 exact accuracy for IMDB-WIKI

(Rothe *et al.*, 2015) dataset. Also (Al-Shannaq & Lamiaa, 2020) used Specific domain transfer learning to train VGGFace network for age estimation. Tingting *et al.* designed a three stage CNN network that consists from three sub CNN networks namely preliminary extraction module, local feature encoding module and recall module for human age estimation (Tingting *et al.*, 2019). Foggia *et al.* developed nine multi-task CNN to recognize age, gender, emotion, and ethnicity. The network architecture is a fusion of three pre-established networks, namely MobileNet, ResNet, and SENet (Foggia *et al.*, 2023). To enhance the precision of age prediction by minimizing overlap of facial features across age ranges, a technique called multi-stage feature constraints learning method has been introduced by (Xia *et al.*, 2020). for face age estimation. This method involves refining the features in three stages by continually updating the feature center for each age range and reducing the distance between each age feature and the corresponding age range's feature center through feature constraint.

3. CONVOLUTIONAL NEURAL NETWORK STRUCTURE

This study utilizes the Soft Stagewise Regression Network (SSR-Net), a convolutional neural network (CNN) developed by (Yang *et al.*, 2018)., as it has demonstrated superior performance in age estimation when compared to other techniques. The SSR-Net contains two separate streams each one is an outstanding convolutional neural network, the inner output among both streams is computed via prediction block and final network output is calculated by the fusion block (see Figure 3b). The pooling size is fixed at 2x2 for all stages. The first stream employs average pooling and the RELU activation function. Tanh activation and max pooling are employed in the second stream. Both pathways implement a convolutional layer with a (3x3) filter, and a fusion block in (Figure 3) employs a (1x1) convolutional layer. The overall parameter count for this network is 40,915. Finally, the soft stage wise regression is carried out.

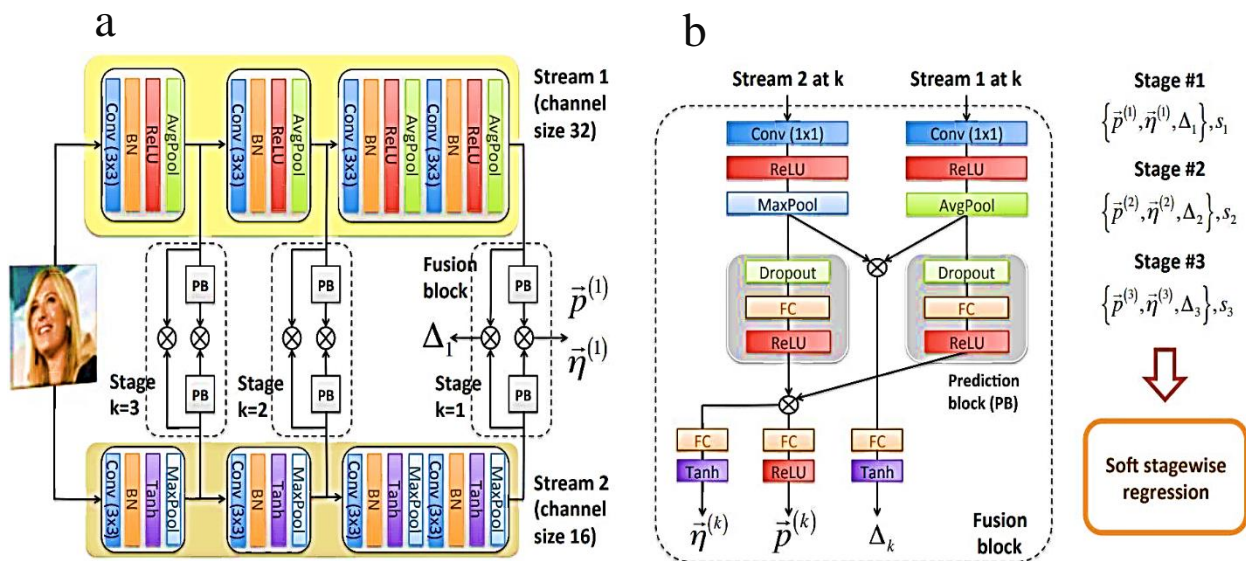


Fig (3): The soft stagewise regression network structure. (a) CNN for feature extraction, (b) Fusion block and regression approach (Yang *et al.*, 2018).

Using different activation functions (RELU and Tanh) and different Pooling approaches (Average and Maximum) in each stream makes the network heterogeneous. Doing so the network can learn different features during the training process and their fusion will improve the performance of the age estimation. The fusion

block is responsible for generating the stage wise output, in this block the features from both streams flow through the 1x1 convolutional layer, activation and pooling to adopt more compact features. Two feature maps are obtained by element wise multiplication and fed to a fully-connected layer followed by Tanh activation

function to obtain the value in a range of (-1, 1). ReLU is used as its activation for obtaining positive values. On the other hand, Tanh is used to allow shifts on both positive and negative sides.

4. THE LOSS FUNCTIONS

The loss function plays a very important role in every deep learning model in network efficiency and accuracy. The network structure cannot determine the classification or regression task; this can only be done by choosing the correct loss function. For instance, if the problem is the classification then a probabilistic loss function such as (categorical cross entropy, Poisson function or KLDivergence class) should be used. On the other hand, if the problem is a regression task, then Mean Absolute Error (MAE) or Mean Squared Error (MSE) loss function should be used. In this paper, all work is focused on loss functions, and a new loss function is developed due to lack of special loss function for regression tasks using convolutional neural networks.

In a neural network model, the loss function is employed to measure the success of the network's performance by converting a set of parameter values into a scalar value. This value indicates how effectively the parameters execute the task for which the network was designed. The objective function, also known as a cost or loss function, is utilized in neural networks to produce the error. Which is the distance between the ground truth label and predicted label of each image sample participated in the training process. The value produced by the loss function is typically called "loss."

4.1 The Mean Squared Error (MSE)

The mean squared error (MSE) of a model measures the average of the squares of errors, which is the average squared difference between the observed values and what is estimated, which can mathematically be represented in (Equation 1). MSE is a risk function that represents the expected value of the squared error loss (Lehmann & Casella, 1998). The fact that MSE is almost always strictly positive (and greater than zero) is due to randomness of weights or because the estimator fails to take into account information that could produce a more accurate result.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^p)^2$$

where, n is total number of items, y_i is the actual label value and y_i^p is the estimated label value.

In MSE the error is the square of the difference between actual and predicted ($y_i - y_i^p$) value. The value of this error increases significantly if the error is greater than one. If the training data contained outliers this will cause a very high value of error which will result in very much final error or very small error that will cause loss vanishing at the ending epochs of the training process. This will make the model trained with MSE give more weight to outliers rather than the important features of the training data. This is the reason why MSE is always almost greater than one, and struggle from minimizing the loss of the network and very difficult to find the global minima. This can be seen graphically in the (Figure 4), which shows the range of prediction for MSE loss functions, which is from zero to infinity.

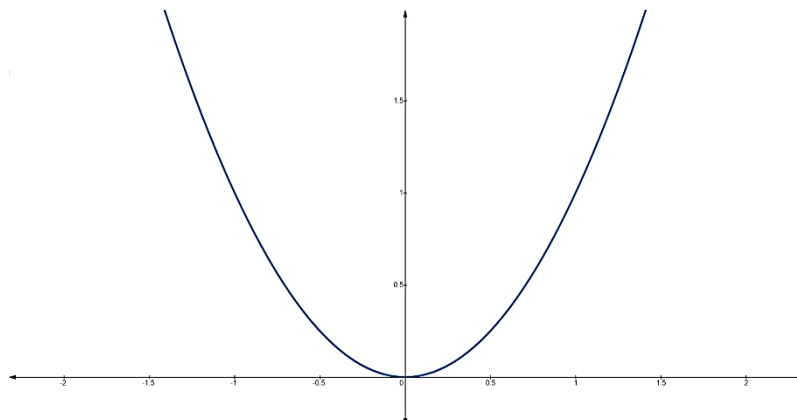


Fig (4):The plot of MSE loss function (Y-axis) vs. (X-axis) the predictions range is between zero and infinity.

The Root Mean Squared Error (RMSE), which is only the square root of MSE to make it

on the same scale as MAE (Mean Absolute Error). The model trained with RMSE loss

function, will be tweaked to minimize the single outlier case at the expense of other common features of the training data, which will lower its overall performance. Because RMSE is also calculated by squaring errors and calculating a mean, it can be heavily influenced by a few predictions that are significantly worse than the MSE. Thus, using the absolute value of both true and target labels and/or calculating the median can provide a better idea of how a model performs on most predictions while excluding the extra influence of unusually poor predictions.

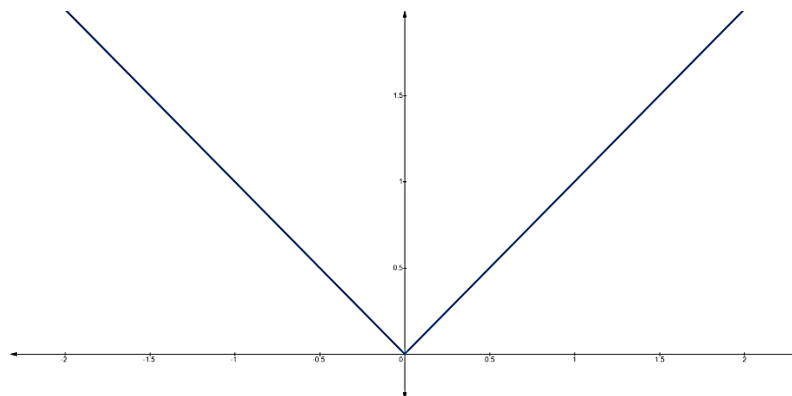


Fig (5): The plot of MAE loss function (Y-axis) vs. (X-axis) the predictions range is between zero and infinity.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y_i^p|$$

where, n is total number of items, y_i is the actual label value and y_i^p is the estimated label value.

One major issue with using MAE loss function for neural nets in particular is that its gradient is constant throughout, implying that the gradient will be large even for small loss values. This is not conducive to learning. To address this, we can employ a dynamic learning rate that decreases as we get closer to the minima. In this case, MSE behaves well and will converge even with a fixed learning rate. The gradient of MSE loss is high for larger loss values and decreases as loss approaches zero, making it more precise at the end of training. It is worth mentioning that, both MSE and MAE ignore the negative losses by squaring and taking the absolute respectively. This cannot give the true distance between the actual and predicted values by forcing the loss to flow in the positive direction only.

4.2 The Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a sum of absolute differences between actual and predicted values as mathematically can be represented in (Equation 2). This loss function is used for regression purposes. Thus, it calculates the average magnitude of errors and ignores their directions. If the directions are considered it will be Mean Bias Error (MBE), which is the sum of the differences between predicted and actual values. The MAE is more robust to outliers and can be graphically represented in (Figure 5). The prediction range is also between zero and infinity.

4.3 The Proposed Golden Mean Squared Error (GMSE)

In this work we propose a new loss function namely Golden Mean Squared Error (GMSE) that can be used to measure the error of the neural nets for regression problems. Which is driven from both MSE (Equation 1) and golden ratio (Equation 3).

$$\text{Golden Ratio} = 1 + \frac{\sqrt{5}}{2}$$

The GMSE loss function is created by implementing MSE in the golden ratio equation and change 1 in (Equation 3) to β as a variable we have:

$$GMSE = \beta + \frac{\sqrt{\sum_{i=1}^n (y_i - y_i^p)^2}}{2n}$$

where, β is the bias, n is total number of items, y_i is the actual label value and y_i^p is the estimated label value.

The GMSE can be seen as a collection of MSE, RMSE and MAE losses. This new loss function focuses on details more than outliers by the role of the square root that works as RMSE. In addition, dividing that error by (2) will reform the error as it is produced by MAE. Thus, the

error is measured in many directions and in different ways which makes it more robust to homogeneous data and can converge much faster that will lead to finding the global minima. The

GMSE can be graphically represented in (Figure 6). That shows the full GMSE loss function, the prediction range is between β and infinity.

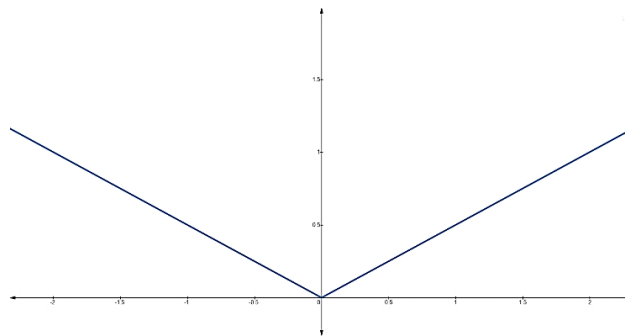


Fig (6):The plot of GMSE loss function where $\beta=0$, (Y-axis) vs. (X-axis) the predictions range is between β and infinity

4.3.1 The role of β

As it can be seen in (Figure 6) the prediction range of GMSE is between zero and infinity, this is due to the role of β in (Equation 4) as it is a part of golden ratio's equation, which shifts the start range from zero to one or negative one. This is very important to neglect some outliers and give the focus of learning on more important details. The β_{\pm} can be set between -1 and 1 which is used to set the base of the learning curve. Hereafter, the β_{\pm} which is indicated as the basis of the function which represents the base of the learning curve, will change the prediction range between zero and infinity to β and infinity.

4.3.2 The role of dividing by (2)

Dividing the calculated error amount by two opens the gate for more features to participate in the learning process. This makes the curve of the GMSE wider than MAE and less steep than the MSE curve that can be graphically seen in (Figure 7). Which plays a very important role in learning of features and finding the best local minima of the network. By squaring the error as MSE does the amount of loss is getting larger this will decrease the chance of approaching the regression line. This loss will work as a momentum facing the network to avoid being converged, the reason behind this is the existence of the square in MSE. Therefore, taking the square root and dividing it by (2) will transform it into a smaller scale. Which will produce a much smaller loss. Eventually as the empirical loss of a model decreases the model performs

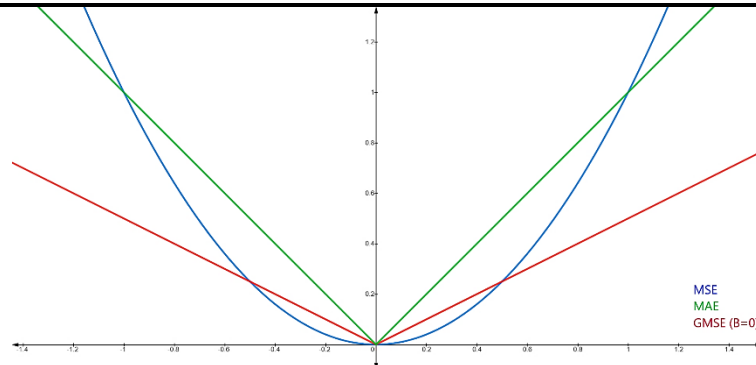
better in terms of accuracy and simplicity. As long as there are alternative methods to reduce loss besides altering the network structure, then simplifying the network can save a significant amount of time and training costs. Therefore, the loss function plays a crucial role in reducing the overall loss of the network.

4.3.3 The Role of Square Root

The square used in MSE will always give a positive value of error, preventing errors from canceling each other out, and it gives more weight to values further away from the target function, emphasizing points where the estimator is poor. Therefore, the square root is used in GMSE to remove the effects of the squaring. Because the square root reverts the squared units to their original dimensions much like variance versus standard deviation. This is a pure mechanism to avoid the loss vanishing phenomena, which commonly occurs in deep networks and prevents the model optimizer from becoming stuck in the local minima pushing it downward to stabilize at the global minima (Mohammed & Tahir, 2020). Thus, in many cases when using MSE as a loss function the network produces a high amount of error value at the beginning stages of the training and a very small amount of loss at the final stages of learning as compared with other losses in (Table 1). The GMSE loss function can give a longer breath to the network due to the usage of the square root.

Table (1): Comparison of Loss Functions.

MSE	MAE	GMSE
Does not contains any parameters.	Does not contains any parameters.	Contains a parameter β that can be used to setting the learning curve. See Equation (4)
Works better for classification task	Specially designed for machine learning approaches.	Specially designed for deep learning and esitimation tasks
Slow in finding global minima	Fast in finding global minima	Much Faster in finding global minima
Produces very large amount of loss at the begaining echpocs of training	Produces large amount of loss at the begaining echpocs of training	Produces very small amount of loss at the begaining echpocs of training

**Fig (7):**The graphical representation of MSE in blue color, MAE in green color and GMSE in red color loss functions.

5. DATASETS

In this paper, two different datasets were used, FG-NET dataset (Lanitis, 2008) and UTKFace (Version 1) (Zhang *et al.*, 2017). The UTKFace is a large-scale face dataset with a large age variance (range from 0 to 116 years old). This dataset contains over 23,000 facial images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. This dataset could be used on a variety of tasks, e.g., face detection, age estimation, age progression/regression, landmark localization, etc. The FG-NET dataset contains one thousand face images from 82 persons with medium variation of lighting, pose, and expression. The age ranges from 0 to 69 (on average, 12 images per person). Although there are much bigger datasets that are very suitable for age estimation purposes, those datasets may include (IMDB-WIKI (Rothe *et al.*, 2015), MORPH (Ricanek & Tesafaye, 2006), and MegaAge-Asian (Zhang *et al.*, 2017)) datasets. Also, there are some other types of datasets such as OU-ISIR gait-based dataset that was used in (Hassan *et al.*, 2018) for human age and gender classification.

The labels of images in UTKFace dataset are included in the image file name and are formatted as [age] is an integer from 0 to 116, indicating the age of the person, [gender] is either 0 (male) or 1 (female), [race] is an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern) and [date&time] is in the format of yyyyymmddHHMMSSFFF, showing the date and time an image was collected to UTKFace. The FG-Net dataset labels are also stored in the file name of the picture and is formatted as [001] first three digits indicating the person Identification Number (ID) followed by letter 'A' this character is used to prevent the conflict with other image files. Finally, the file name is ended with two digits indicating the age of the person.

6. RESULTS & DISCUSSION

The SSR-Net CNN is applied to UKTFace and FG-NET datasets using the proposed loss function GMSE. For the purpose of comparison, the same CNN is implemented on the same datasets using MSE and MAE loss functions. The network has been trained on Kaggle, the data science and machine learning platform using Graphical Processing Unit (GPU). All the code is

written in python programming language with the assistance of TensorFlow and Keras libraries. In order to maximize the performance of the CNN, the hyperparameters shown in (Table 2) are selected. In the training process, each dataset

was divided into two parts 80 percent for training and the remaining 20 percent was used for validation process. The results for training and testing phases are demonstrated in the following subsections.

Table (2): Hyperparameter Initialization.

Hyperparameter	Initialization
Input size	64x64x3
Learning rate	0.001
Optimizer	Adam
Dropout	0.2
Batch size	64
Epochs	250
GMSE Loss (β)	0

6.1 Training results

Figures (8 and 9) show the three loss functions versus epoch number for the two datasets, UKTFace and FG-NET. The red curve represents the proposed loss function GMSE, the green curve represents the MAE loss function and the blue curve represents the MSE loss function. The best results achieved by GMSE loss function for both datasets. For UKTFace dataset, the amounts of loss achieved at the first epoch by MSE, MAE and GMSE are 474.9, 12.76 and 7.91 respectively. These amounts of loss drop significantly and reach 51.34, 4.8 and 3.22 respectively for MSE, MAE and GMSE at epoch 25. These losses continue to drop with

increasing the number of epochs. At epoch 250, GMSE loss is dramatically reduced to 1.58, while MAE loss is just above 2 and MSE loss is 6.56. The loss values of GMSE and MSE indicate that the loss value that can be achieved by GMSE at epoch 25 is better than that which can be achieved by MSE, even at epoch 250. This result approves that the training of SSR-Net CNN is faster when GMSE loss function is used compared to the use of MSE. This is due to the mathematical operation of square in MSE function. In addition, the GMSE and MAE curves look smoother than the MSE curve, which indicates the high fluctuation (instability) in the MSE loss values during the training phase.

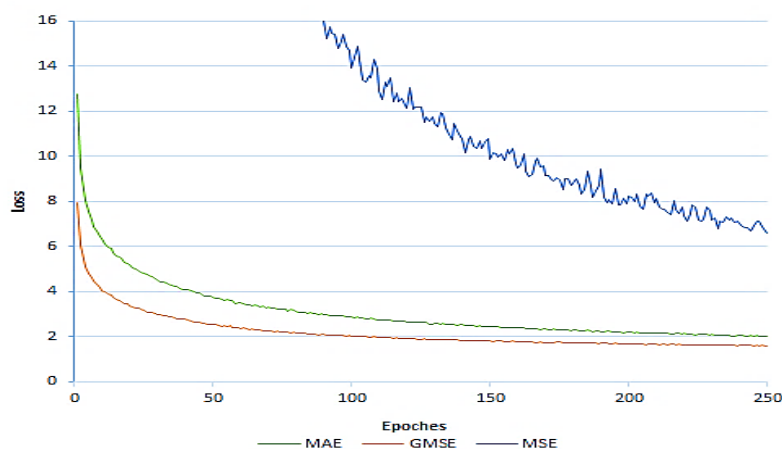


Fig (8):The training loss of UTKFace Dataset over 250 epochs. The GMSE demonstrates superior performance compared to other types of loss functions.

For FG-NET dataset, the same scenario can be seen in (Figure 9). At the first epoch, the loss values for GMSE, MAE and MSE are 8.92, 15.54 and 496.31 respectively. These loss values drop with increasing the number of epochs. At epoch 250, the loss values of GMSE, MAE and MSE become 1.34, 1.72 and 6.58 respectively.

The similarity in the changes of loss values with different datasets approves that the hyperparameters and the CNN structures are adaptable to different datasets. The (Table 3), shows details of a comprehensive comparison of the loss functions for the two datasets.

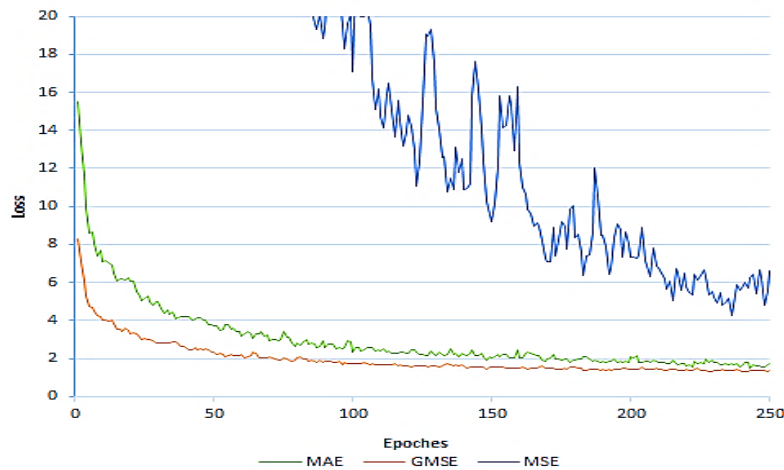


Fig (9):The training loss of FG-NET dataset over 250 epochs.

Table (3): Losses at different epochs for UTKFace and FG-NET Datasets. The results are the average of several runs.

Dataset	Epoch No.	MSE	MAE	GMSE
UTKFace	1	473.92	12.76	7.91
	25	51.34	4.80	3.15
	100	13.91	2.85	2.00
	250	6.56	2.02	1.58
FG-NET	1	496.31	15.54	8.29
	25	63.65	5.15	3.05
	100	17.09	2.35	1.72
	250	6.58	1.72	1.34

In order to show the efficiency of the proposed loss function, it is compared to some of the benchmark results for age estimation that have used the same datasets, FG-NET and UKTFace. The comparison is based on the differences achieved between the actual and estimated age. (Tables 4 and 5) show the comparisons. According to these tables, the proposed loss function with the SSR-Net CNN outperforms the previous works for both,

accuracy and speed. This is due to two reasons. First, the proposed loss function is more effective compared to those used in previous works. Second, in previous works, several techniques and approaches were used and this may require more computation time, while in the proposed method, only a loss function is used which can achieve higher accuracy with less computation time for training process.

Table (4): Comparison with the state-of-the-art methods on UTKFace dataset for age estimation task

Loss Function	Method	Difference in Years
MAE	CORAL (Cao et al. 2020)	5.39
MAE	Specific Domain (Al-Shannaq and Lamiaa 2020)	4.86
MAE	Randomized Bins (Berg et al, 2021)	4.55
MAE	TResNet-S (Yoshimura and Ogata 2020)	4.49
MAE	SSR-Net (Yang et al. 2018)	2.02
GMSE (Ours)	SSR-Net CNN + GMSE Loss	1.58

Table (5): Comparison with the state-of-the-art methods on FG-NET dataset for age estimation task.

Loss Function	Method	Difference in Years
MAE	Specific Domain (Al-Shannaq and Lamiaa 2020)	3.44
MAE	C3AE (Zhang et al, 2019)	2.95
MAE	Deep Age Estimator (Hu et al, 2017)	2.80
MAE	Mean Variance Loss (pan et al, 2018)	2.68
MAE	SSR-Net (Yang et al. 2018)	1.72
GMSE (Ours)	SSR-Net CNN + GMSE Loss	1.34

6.2 Testing results

The testing accuracy indicates how the model will perform at the production time with unseen data. A large number of samples from both dataset is taken and tested with the trained models. Their performance is evaluated using the accuracy metric. The accuracy takes the absolute value of the distances between true and predicted labels and subtracts the output from 100. The accuracy metric can be mathematically represented in the following equation. $acc = 100 - |y_i - y_i^p|$ (5)

where y_i is the actual label and y_i^p is the predicted label.

In the testing phase, the GMSE also outperformed MSE and MAE by achieving 97.21% percent in accuracy for UTKFace dataset and 97.53% percent for FG-NET dataset. Meanwhile, the MAE couldn't achieve 97% percent accuracy for any datasets. Moreover, the MSE could achieve 97.28% only for the FG-NET dataset. (Table 7) shows the testing accuracy for all three loss functions according to each dataset; it can be seen that MSE and MAE perform better for small datasets than large datasets. However, GMSE can achieve similar results for large and small datasets.


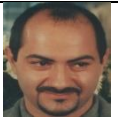





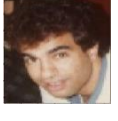


Table (6): Testing Accuracy for UTKFace and FG-NET Dataset

Dataset	Loss Function	Accuracy
UTKFace	MSE	95.82
	MAE	96.32
	GMSE	97.21
FG-NET	MSE	97.28
	MAE	96.96
	GMSE	97.53

(Table 7), depicts the actual and estimated ages for some samples taken arbitrarily from the two datasets, UTKFace and FG-NET using different loss functions. The table shows that the estimated

ages achieved by GMSE are closer to the actual ages compared to what have achieved by MAE and MSE loss function.

Table (7): Comparison of the testing results for five random samples from UTKFace and FG-NET dataset using SSR-NET structure with three different loss functions MSE, MAE and GMSE.

UTKFace					FG-NET				
<i>Input Image</i>	<i>Actual Age</i>	<i>MSE</i>	<i>MAE</i>	<i>GMSE</i>	<i>Input Image</i>	<i>Actual Age</i>	<i>MSE</i>	<i>MAE</i>	<i>GMSE</i>
	29	28.28	22.01	29.5		30	30.99	30.02	30.56
	82	87.96	80.43	78.66		13	12.71	12.72	12.65
	2	1.29	1.01	2.95		3	3.93	4.45	3.20
	32	30.72	28.92	31.08		23	22.32	25.86	22.86
	10	8.49	9.10	10.25		7	6.08	9.02	7.81

The experiments also show that, MSE performs better than MAE at test time for small datasets such as FG-NET dataset. In contrast, the MAE is better than MSE for large datasets such as unconstraint UTKFace dataset. Meanwhile the proposed GMSE loss function performs better than both loss functions (MSE and MAE) for small and large datasets.

7. CONCLUSIONS

In this paper, the age estimation task was undertaken, the influence of loss functions for age estimation tasks in particular was studied. A new loss function GMSE was proposed to accurately compute the distance between the actual age label value and the estimated age label value. The GMSE function was obtained by merging the golden ratio and MSE loss functions. By defining three new terms (bias, square root and division by two) in the MSE function. The differences between the main loss functions for regression problems were observed. We have seen that GMSE loss function can compute the error more precisely than the state-of-the-art loss functions such as MAE and MSE for regression tasks. The results have shown that GMSE outperformed MAE and MSE for large and small

datasets. We have obtained an error of 1.58 for the UTKFace dataset using the GMSE loss function which is about 44% better than MAE loss function. The results are promising and competitive with the existing loss functions. In the upcoming years, the task of training on extensive datasets like IMDB-WIKI (Rothe et al., 2015), MORPH (Ricanek & Tesafaye, 2006), and MegaAge-Asian (Zhang et al., 2017), along with incorporating gait-based datasets such as OU-ISIR for age estimation, will pose a significant challenge. This challenge arises from the need to predict a person's age based on their gait, derived from their body movements.

REFERENCES

- Agbo-Ajala, O., & Viriri, S. (2020). Deeply Learned Classifiers for Age and Gender Predictions of Unfiltered Faces. *The Scientific World Journal*, 2020, 1-12.
- Al-azzawi, D. S. (2019). Human Age and Gender Prediction Using Deep Multi-Task Convolutional Neural Network. *Journal of Southwest Jiaotong University*, 54(4), 1–11.
- Al-Shannaq, A., & Lamiaa, E. (2020). Age Estimation Using Specific Domain Transfer Learning. *Jordanian Journal of Computers and*

- Information Technology (JJCIT), 6(2), 122–139.
- Berg, A., Oskarsson, M., & O'Connor, M. (2021). Deep Ordinal Regression with Label Diversity. 2020 25th International Conference on Pattern Recognition (ICPR), vol. 2, 2740–2747.
- Bui, T., Ribeiro, L., Ponti, M., & Collomosse, J. (2018). Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. *Computers & Graphics*, 71, 77–87.
- Dhimar, T., & Mistree, K. (2016, March). Feature extraction for facial age estimation: A survey. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 2243-2248). IEEE.
- Dornaika, F., Bekhouche, S. E., & Arganda-Carreras, I. (2020). Robust regression with deep CNNs for facial age estimation: An empirical study. *Expert Systems with Applications*, 141, 112942-112948.
- Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12), 2170–2179.
- Foggia, P., Greco, A., Saggese, A., & Vento, M. (2023). Multi-Task Learning on the Edge for Effective Gender, Age, Ethnicity and Emotion Recognition. *Engineering Applications of Artificial Intelligence*, 118(1), Article ID 105651.
- Guo, G., Mu, G., Fu, Y., Dyer, C., & Huang, T. (2009). A study on automatic age estimation using a large database. In *IEEE 12th International Conference on Computer Vision* (pp. 1986-1991).
- Gupta, S. K., & Nain, N. (2023). Single attribute and multi attribute facial gender and age estimation. *Multimedia Tools and Applications*, 82(1), 1289-1311.
- Han, H., Otto, C., & Jain, A. K. (2013, June). Age estimation from face images: Human vs. machine performance. In 2013 international conference on biometrics (ICB) (pp. 1-8). IEEE.
- Hassan, O. M. S., Abdulazeez, A. M., & Tiryaki, V. M. (2018). "Deeply learned classifiers for age and gender predictions of unfiltered faces." *The Scientific World Journal*, 2020 (1289408).
- Hu, Z., Wen, Y., Wang, J., et al. (2017). Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7), 3087–3097.
- Jun, T. J., Eom, Y., Kim, D., Kim, C., Park, J. H., Nguyen, H. M., ... & Kim, D. (2021). TRK-CNN: transferable ranking-CNN for image classification of glaucoma, glaucoma suspect, and normal eyes. *Expert Systems with Applications*, 182, 115211.
- Kendall, A., & Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. In *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017* (pp. 6555–6564).
- Lanitis, A. (2008). Comparative Evaluation of Automatic Age-Progression Methodologies. *EURASIP Journal on Advances in Signal Processing*, 2008(1), Article ID 239480.
- Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer.
- Liu, N., Zhang, F., & Duan, F. (2020). Facial Age Estimation Using a Multi-Task Network Combining Classification and Regression. *IEEE Access*, 8(14), 92441–92451.
- Lu, X., Ma, C., Ni, B., et al. (2018). Deep regression tracking with shrinkage loss. *Proceedings of the European conference on computer vision (ECCV)*, 353-369.
- Mohammed, A. I., & Tahir, A. A. (2020). A New Optimizer for Image Classification using Wide ResNet (WRN). *Academic Journal of Nawroz University*, 9(4), 1-13.
- Pan, H., Han, H., Shan, S., & Chen, X. (2018). Mean-Variance Loss for Deep Age Estimation from a Face. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5285-5294).
- Ricanek Jr., K., & Tesafaye, T. (2006). MORPH: A longitudinal image Age-progression, of normal adult. In *Proc. 7th Int. Conf. Autom. Face Gesture Recognit* (pp. 0–4).
- Rothe, R., Timofte, R., & L. Van Gool. (2015). DEX: Deep EXpectation of Apparent Age from a Single Image. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 10-15).
- Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., & Yuille, A. L. (2018). Deep regression forests for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2304-2313).
- Taheri, S., & Toygar, Ö. (2019). On the Use of Dag-CNN Architecture for Age Estimation with

- Multi-Stage Features Fusion. *Neurocomputing*, 329, 300–310.
- Tahir, A. A. (2012). Integrating artificial neural network and classical methods for unsupervised classification of optical remote sensing data. *EURASIP Journal on Advances in Signal Processing*, 2012, 1-12.
- Tingting, Y., Junqian, W., Lintai, W., & X.Yong. (2019). Three-stage network for age estimation. *CAAI Transactions on Intelligence Technology*, 4(2), 122-126.
- Toshev, A., & Szegedy, C. (2014). DeepPose: Human Pose Estimation via Deep Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1653-1660.
- Xia, M., Zhang, X., Liu, W., Weng, L., & Xu. Y. (2020). Multi-Stage Feature Constraints Learning for Age Estimation. *IEEE Transactions on Information Forensics and Security*, 15, 2417-2428.
- Yang, S., Lin, W., Lin, G., Jiang, Q., & Liu, Z. (2021). Progressive Self-Guided Loss for Salient Object Detection. *IEEE Transactions on Image Processing*, 30(4), 8426-8438.
- Yang, T. Y., Huang, Y. H., Lin, Y. Y., Hsiu, P. C., & Chuang, Y. Y. (2018). SSR-NET: A compact soft stagewise regression network for age estimation. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 1078–1084.
- Zhang, C., Liu, S., Xu, X., & Zhu, C. (2019). C3AE: Exploring the limits of compact model for age estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12587-12596).
- Zhang, K., Gao, C., Guo, L., et al. (2017). Age Group and Gender Estimation in the Wild with Deep RoR Architecture. *IEEE Access*, 5, 22492–22503.
- Zhang, Y., Liu, L., Li, C., & Loy, C. C. (2017). Quantifying Facial Age by Posterior of Age Comparisons. *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 126-145.
- Zhang, Z., Song, Y., & Qi, H. (2017). Age Progression / Regression by Conditional Adversarial Autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5810–5818).