Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

An Optimized Deep Transfer Learning Framework for Robust Human Action Recognition in Challenging Environments

Muhammad Adnan Javed
Lecturer, HITEC University Taxila Cantt Email: adnanjaved2122@gmail.com

Ayesha Riaz
Student, HITEC University Taxila Cantt Email: ayeshariaz4144@gmail.com

Muhammad Abdullah
Student, HITEC University Taxila Cantt Email: abdullahkamran4@gmail.com

Muhammad Abdul Moeez Bashir
Student, HITEC University Taxila Cantt Email: bashirmoeez@gmail.com

Irfan Haider
Student, HITEC University Taxila Cantt Email: irfan.haider@hitecuni.edu.pk

Hanzala Rehman
Student, HITEC University Taxila Cantt Email: rehmanhanzala203@gmail.com

ABSTRACT

The most important research area in computer vision is action recognition (AR) based on AI reasoning. During the past few years, big breakthroughs have been witnessed in this field, also the interest of research in this field is evolving like understanding of actions and scenes, studying the human joints, and human posture recognition. The precision of AR has been raised due to the growing of learning-based AI. Even though various innovations are being witnessed in AI technology, but still there exist quite a few challenges also in this field. This field requires large datasets and corresponding labels for their learning-based algorithms. Human action recognition (HAR) is rapidly gaining popularity in the field of pattern identification due to its unobtrusive features acquired from a distance and the fact that, unlike some other biometric applications like iris and fingerprint detection, it does not require the participation of individuals. Robotics, location estimation, sports analysis, pedestrian detection, human-computer interaction, video games, and video surveillance, are some examples of real-world applications where HAR is applied. Although different research approaches for action recognition have been presented, these methods still face obstacles like different viewpoints, environment conditions, and temporal variations, etc. The main purpose of this study is to deal with these addressed challenges. The suggested technique consists of a few core stages i.e. frames initial preprocessing, train using transfer learning concept, features extraction, fusing extracted features, features optimization, and classification. The first step is image

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

preprocessing, in which the image dataset will be normalized by using color transformation and resizing. Pre-trained deep networks namely InceptionV3 and Resnet50 are used to extract deep features and then fused these feature vectors. Optimization based on IWOA is performed on fused vector and finally, classifiers are used for classification. Four publicly accessible datasets such as UCFT youtube, wizman, Ixmas, and UCF Sports are used to validate the HAR method and attained an accuracy of 100%, 99.9%, 99.1%, and 100% respectively.

Keywords: Optimized, Deep Transfer, Learning Framework, Robust, Human Action Recognition, Challenging Environments.

Introduction

Human Action Recognition (HAR) has evolved into a pivotal area of research within computer vision, driven by its wide-ranging applications such as video surveillance, healthcare monitoring, smart homes, and human-computer interaction (Poppe, 2010). The ability to accurately interpret human actions from video data enables automated systems to understand and respond to dynamic behaviors in real-world environments. Despite its significance, HAR remains a challenging task due to factors such as viewpoint variations, occlusions, cluttered backgrounds, and intra-class similarities, which complicate reliable action identification in uncontrolled settings (Chaquet et al., 2013).

Traditional HAR methodologies have predominantly relied on handcrafted feature descriptors—including Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and Local Binary Patterns (LBP)—combined with conventional classifiers. While these approaches have achieved reasonable success in constrained environments, they often lack robustness in the face of real-world variability due to their dependence on manually engineered features and sensitivity to environmental changes (Turaga et al., 2008). The advent of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized HAR by enabling automatic extraction of hierarchical spatio-temporal features directly from raw video data. Pre-trained models like ResNet, Inception, and VGG, when adapted through transfer learning, have demonstrated superior performance in capturing discriminative patterns across diverse action classes (Simonyan & Zisserman, 2014; He et al., 2016). However, the high dimensionality of deep features and the presence of redundant information often lead to increased computational complexity and potential overfitting, underscoring the need for effective feature selection and optimization techniques.

In response to these challenges, this paper introduces an optimized deep transfer learning framework designed to enhance the robustness and efficiency of HAR in complex scenarios. The proposed approach leverages a pre-trained ResNet50 model, fine-tuned via deep transfer learning to extract rich spatial features from video frames. To address feature redundancy and improve computational efficiency, an Improved Whale Optimization Algorithm (IWOA) is employed for optimal feature selection. The refined feature set is subsequently classified using multiple classifiers, including Cubic SVM and Linear Discriminant, to achieve high recognition accuracy. Experimental evaluations on five benchmark datasets—KTH, UCF Sports,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Weizmann, UCF YouTube, and IXMAS—validate the efficacy of the proposed framework, demonstrating near-perfect recognition rates and reduced computational overhead in challenging environments.

Problem Statement

Human action recognition involves multiple interdependent stages, including video preprocessing, foreground segmentation, feature extraction, and classification. A core challenge lies in the accurate segmentation of foreground regions within video sequences, which is adversely affected by complex backgrounds, variable lighting conditions, dynamic human motion, and viewpoint changes. These issues are particularly problematic when distinguishing between visually similar actions, such as running and jogging. Furthermore, factors like occlusions, shadows, diverse clothing, and background clutter complicate both segmentation and subsequent feature representation. The extraction and selection of discriminative features also present a significant hurdle, as irrelevant or redundant features can degrade recognition performance while increasing computational demands. Consequently, existing HAR systems often fail to achieve an optimal balance between accuracy and efficiency, especially in real-world, unconstrained environments where computational resources may be limited. Therefore, this research seeks to overcome these limitations by developing a computationally efficient and robust framework that integrates deep transfer learning with intelligent feature optimization to enable reliable human action recognition in challenging scenarios

Literature Review

The field of Human Action Recognition (HAR) has undergone a significant transformation, evolving from traditional handcrafted feature-based approaches to sophisticated deep learning frameworks. Early research predominantly relied on extracting and fusing manually engineered features. For example, Althloothi et al. (2014) introduced a mechanism combining multi-feature extraction with multi-kernel learning for RGB video sequences, fusing shape and kinematic features into a new vector classified by a multi-class SVM. Similarly, other studies focused on local descriptors; Yuan et al. (2016) developed a discriminative parameterized representation, while Hsieh et al. (2016) integrated spatial-temporal HOG with key-point histograms. These methods, along with techniques employing SIFT within Bag-of-Words models (Moussa et al., 2015) or Discrete Cosine Transforms (Ahmad et al., 2015), achieved commendable accuracy on controlled datasets. However, their reliance on manual feature engineering made them inherently sensitive to real-world challenges like occlusion, viewpoint variation, and cluttered backgrounds. The typical pipeline involved sequential stages of pre-processing, foreground segmentation, handcrafted feature extraction, and classification (Liu et al., 2010; Barbu, 2014), a process that struggled with generalization in unconstrained environments.

The advent of deep learning marked a paradigm shift, with Convolutional Neural Networks (CNNs) enabling the automatic learning of hierarchical spatio-temporal features directly from raw video data. A landmark study by Ji et al. (2013) pioneered the use of 3D CNNs to capture temporal information from video volumes. This was significantly advanced by the two-stream CNN architecture (Simonyan & Zisserman,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

2014), which processed spatial (RGB frames) and temporal (dense optical flow) information in separate streams, dramatically improving recognition rates. Subsequent research explored more specialized and hybrid deep architectures. Chenarlogh et al. (2019) investigated various CNN stream configurations for scenarios with limited data, while Bilal et al. (2021) designed a spatiotemporal framework utilizing transfer learning, deep autoencoders, and LSTM networks to handle long videos with overlapping actions. Hybrid models that fuse different feature types also proved effective; Jaouedi et al. (2020) combined traditional HOG features with deep-learned features, and Muhammad et al. (2021) integrated Dilated CNNs with Bi-directional LSTM networks enhanced by an attention mechanism. These deep learning approaches consistently delivered state-of-the-art performance on challenging benchmarks like UCF101 and HMDB51, demonstrating a superior capacity to manage complex dynamics and environmental noise compared to earlier methods.

Despite their power, deep learning models extract high-dimensional feature vectors that often contain redundant information, leading to increased computational complexity and potential overfitting. This has necessitated the development of effective feature selection and optimization strategies. Beyond traditional dimensionality reduction techniques like PCA, recent work has incorporated metaheuristic optimization algorithms. Guha et al. (2020) proposed a Cooperative Genetic Algorithm (CGA) to select the most discriminative subset from a pool of handcrafted and deep features, improving accuracy while reducing the feature vector size. Similarly, nature-inspired algorithms like the Whale Optimization Algorithm (WOA) have been adapted for feature selection in HAR. The WOA, which mimics the bubble-net hunting behavior of humpback whales, and its improved variants (IWOA) are used to navigate the high-dimensional feature space, identifying an optimal subset that maintains high discriminative power while significantly lowering computational load for both training and inference.

Table 2.1 Summary of HAR Methods and Datasets

Reference	Year	Methods	Datasets	Accuracy (%)
[156]	2021	Gated RNN, Feature Extraction	UCF Sports	96.8
[157]	2021	Deep Learning, Fusion of Global Avg Pool and FC layers.	KTH,UT-Interaction, YouTube,UCF Sports, and IXMAS	96.6,96.7,100, 99.7, and 89.6Respectively
[158]	2021	New Residual Network, Feature Fusion, and GAP	CAVIAR,UCF101, UCF11, UT interaction	100, 97.3, 99.5, and 99.9 Respectively
[159]	2021	Novel Pose Descriptor, SVM	HCA,CASIA, UCF11, and UT interaction	88.72, 98, 99, and 96.4 Respectively
[131]	2021	Fusion of Extracted Features, Selected Features M-SVM	UCF YouTube, UCF Sports, Weizmann, and KTH	94.5, 99.3, 97.9, and 100 Respectively

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The effectiveness of HAR systems is rigorously evaluated on standardized public datasets using a comprehensive set of performance metrics. Common benchmarks include the KTH dataset (Schuldt et al., 2004) with six basic actions, the Weizmann dataset (Gorelick et al., 2007) with ten actions, the UCF Sports dataset (Rodriguez et al., 2008) focused on athletic movements, the large-scale UCF101 (Soomro et al., 2012), and the multi-view IXMAS dataset (Weinland et al., 2006). System performance is quantified using metrics such as Accuracy, Precision, Recall (Sensitivity), Specificity, and the derived False Positive and False Negative Rates. A summary of contemporary methods and their reported accuracies on these datasets, as shown in Table 2.1, reveals that modern frameworks employing deep feature fusion and intelligent optimization consistently achieve accuracies above 90%, often approaching or reaching 100% on well-defined datasets. This trajectory underscores the field's continuous progress toward building more robust, efficient, and generalizable human action recognition systems capable of operating in complex, real-world settings. Performance measures of the HAR technique in terms of Sensitivity, Accuracy (ACC), Specificity, Precision, FNR, and FPR are shown in

Table 2.2.
Table 2.2 Performance Measure for HAR

S. No	Performance Measure	Formulas
01	Accuracy (ACC)	$\frac{TN + TP}{TN + TP + FP + FN}$
02	Specificity	$\frac{TN}{FP + TN}$
03	Sensitivity	$\frac{TP}{FN + TP}$
04	Precision	$\frac{TP}{FP + TP}$
05	FPR	$1 - Specificity$

Proposed Methodology

The proposed framework for robust Human Action Recognition (HAR) in challenging environments follows a systematic pipeline designed to enhance feature discriminability and computational efficiency. The overarching architecture, illustrated in Figure 3.1, integrates deep transfer learning with an intelligent feature optimization mechanism. The process begins with the extraction of video frames, which undergo initial preprocessing to normalize illumination and reduce noise. A pre-trained ResNet50 model, fine-tuned via deep transfer learning, is then employed to extract rich, high-dimensional spatial features from these frames. To address the inherent redundancy in deep features and improve model efficiency, an Improved

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Whale Optimization Algorithm (IWOA) is applied to select the most salient and discriminative feature subset. Finally, this optimized feature vector is classified using multiple classifiers, including Cubic SVM and Linear Discriminant, to achieve high recognition accuracy. This methodology aims to balance high performance with computational practicality for real-world deployment.

The initial stage involves critical video preprocessing to ensure input quality. Given that raw video sequences are susceptible to variations in lighting, complex backgrounds, and noise, preprocessing is essential for robust feature extraction. Standard techniques including color normalization, histogram equalization for contrast enhancement, and noise reduction filters are applied to each extracted frame. This step standardizes the input, mitigating environmental artifacts that could otherwise degrade the performance of subsequent deep feature extraction, particularly in unconstrained surveillance or real-world scenarios.

Feature extraction is performed using a deep Convolutional Neural Network (CNN), specifically the ResNet50 architecture, leveraged through transfer learning. CNNs are exceptionally suited for this task due to their hierarchical learning capability, which automates the extraction of complex spatial features from raw pixels. The ResNet50 model, pre-trained on the large-scale ImageNet dataset, provides a powerful foundational feature extractor. To adapt this model to the specific domain of human actions, transfer learning is applied. The final fully connected layers of ResNet50 are replaced and the model is fine-tuned on the target HAR datasets. This process allows the network to retain generalized feature knowledge from ImageNet while learning discriminative patterns specific to human actions, thereby improving performance especially when training data is limited. The feature vector is extracted from the penultimate layer of the fine-tuned network.

A core innovation of the proposed framework is the application of the Improved Whale Optimization Algorithm (IWOA) for feature selection. While deep networks extract powerful features, the resulting high-dimensional vectors often contain redundant or irrelevant information, increasing computational load and risk of overfitting. The IWOA, an enhanced variant of the metaheuristic Whale Optimization Algorithm (WOA), is employed to identify an optimal subset of features. The standard WOA simulates the bubble-net hunting behavior of humpback whales through three phases: encircling prey, spiral updating, and random search for prey. The IWOA introduces key improvements to overcome WOA's limitations, such as premature convergence and imbalanced exploration-exploitation. These include an adaptive search-enclosing technique that dynamically adjusts the probability of exploration versus exploitation across iterations, and a "jumping behavior" mechanism that helps the algorithm escape local optima by randomly perturbing whale positions. The IWOA evaluates the fitness of different feature subsets (e.g., based on classifier performance) and iteratively converges on the combination that yields the highest discriminative power with minimal dimensionality. This optimization yields a compact, robust feature vector.

The final stage is classification, where the optimized feature vector is fed into a suite of classifiers to determine the action label. To ensure comprehensive evaluation and identify the best-performing model for the refined features, multiple classifiers are tested, including Fine/Medium/Coarse Trees, Linear/Quadratic/Cubic Support Vector

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Machines (SVM), Linear/Quadratic Discriminant Analysis, and Naive Bayes variants. The classifier achieving the highest accuracy and efficiency on the validation set is selected for the final model. This multi-stage methodology—encompassing sophisticated preprocessing, deep transfer learning-based feature extraction, IWOA-driven feature optimization, and rigorous classifier evaluation—constitutes a comprehensive and optimized framework for accurate and efficient human action recognition.

Experimental Results

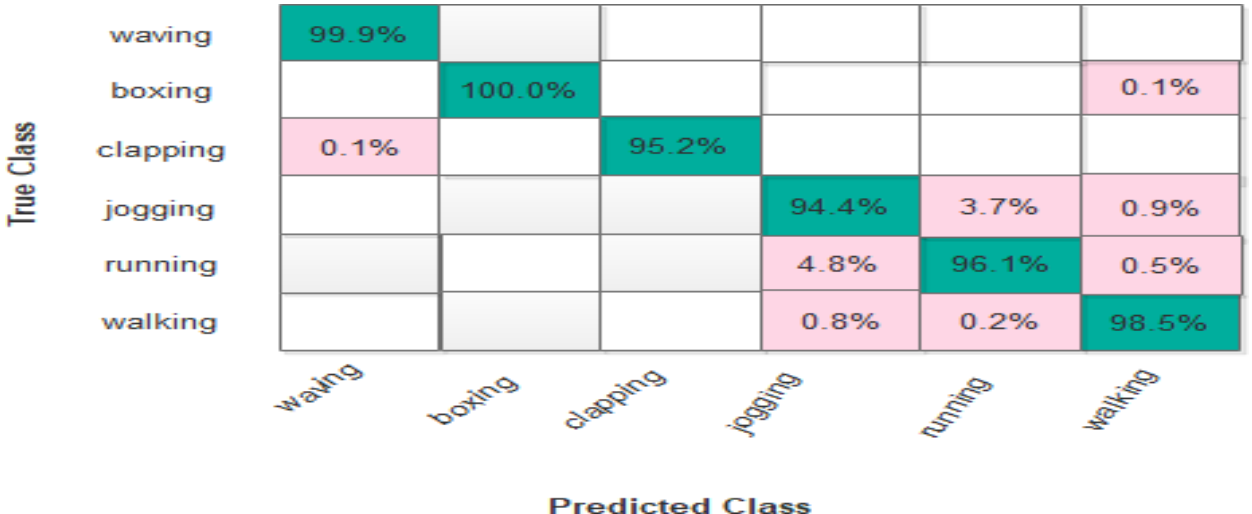
This section comprises a full discussion of the results and analyses. The suggested model is tested on Five different datasets namely, (i) KTH, (ii) UCF Sports, (iii) Weizman, (iv), UCFT Youtube (v) IXMAS dataset. 10 folds cross-validation is used on all 5 datasets. Ten different classifiers, including Fine Tree, Medium Tree, Coarse Tree, Linear Discri Quadratic SVM, minant, Linear SVM, Quadratic SVM, Cubic SVM, Gaussian Naïve Bayes, Quadratic Discriminant and Kernel Naïve Bayes were used for result collection. All experiments are accomplished on MATLAB 2020a, on PC having specification, Core i7 with 8 GB of DDR4 RAM.

Results Proposed Methodology

In this experiment, we used the KTH dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Cubic SVM classifier achieved the highest accuracy of 98 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 98.15, 98.05, 0.005, and 1.0 being calculated, respectively. The confusion matrix for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Cubic SVM classifier is 156.83 seconds. Furthermore, the shortest calculation time is 24.24 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 1 Feature testing using resnet50 for KTH dataset

CLASSIFIER	TPR	PPV	F1-Score	FPR	AUC	ACC %	Time
Fine tree	75.26	76.23	75.74	0.048	0.911	75.40	49.44
Medium tree	65.31	65	65.15	0.071	0.860	64.90	37.12
Coarse tree	47.86	43.85	45.77	0.115	0.870	48.90	24.24
Linear discriminant	97.2	97.33	97.26	0.003	0.990	97.20	29.73



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Linear svm	94.58	94.93	94.75	0.010	0.996	94.50	169.48
Quadratic svm	97.58	97.8	97.69	0.001	1.000	97.60	173.21
Cubic svm	97.96	98.15	98.05	0.005	1.000	98.00	156.83
Gaussian naive bayes	60.33	62	61.15	0.053	0.800	59.40	39.67
Quadratic discriminant	89.86	94.85	92.29	0.020	0.941	90.70	42.49
Kernel naïve bayes	73.56	72.88	73.22	0.056	0.868	72.20	5062.70

Figure 4.1:Cubic SVM classifier‘s confusion matrix on KTH dataset using Resnet50 deep features.

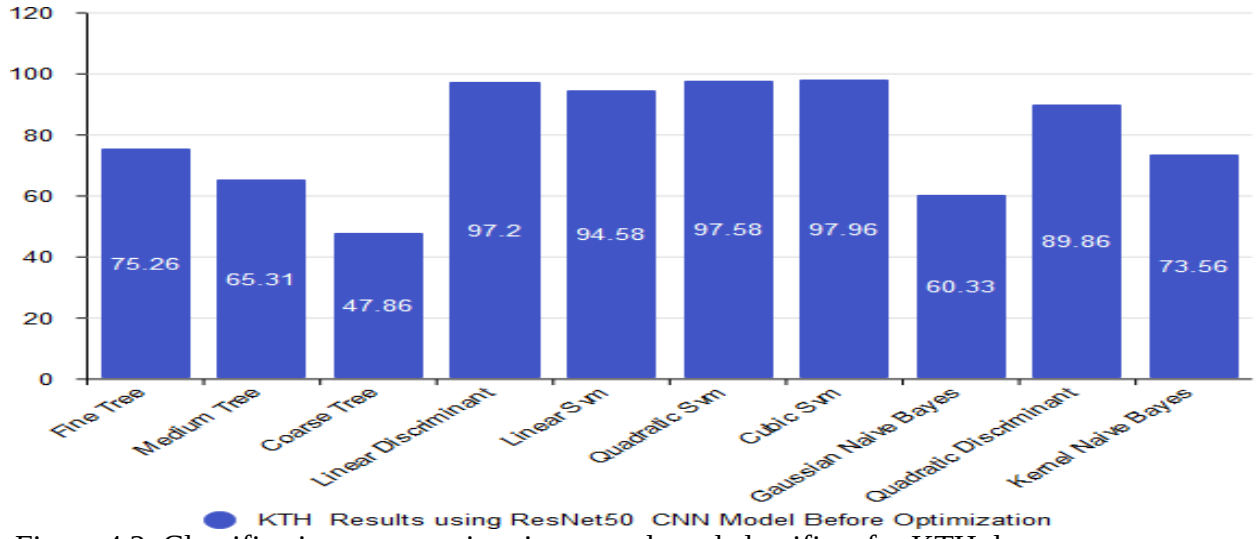


Figure 4.2: Classification computation time on selected classifiers for KTH dataset.

Table 4. 2: Feature testing using resnet50 for UCF Sports dataset

Classifier	TPR	PPV	F1-Score	FPR	AUC	ACC%	Time
Fine tree	68.13	72.35	70.18	0.016	0.943	81.50	135.14
Medium tree	48.59	51.30	49.91	0.030	0.892	68.50	27.76
Coarse tree	31.81	25.23	28.14	0.043	0.730	52.40	22.40
Linear discriminant	100.00	99.94	99.97	0.000	1.000	100.00	30.13
Linear svm	99.47	99.82	99.64	0.000	1.000	99.80	183.53
Quadratic svm	99.89	99.90	99.89	0.000	1.000	99.90	209.24
Cubic svm	99.89	99.90	99.89	0.000	1.000	99.90	201.60
Gaussian naive bayes	84.19	79.30	81.67	0.014	0.941	81.40	119.17
Quadratic discriminant	83.30	81.40	82.34	0.042	0.700	87.90	283.70
Kernel naive							

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

bayes	92.30	87.03	89.59	0.007	0.898	90.00	4003.90

In this experiment, we used the UCF Sports dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Linear Discriminant classifier achieved the highest accuracy of 100 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 99.94, 99.97, 0.00, and 1.0 being calculated, respectively. The confusion matrix for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Linear Discriminant classifier is 30.13 seconds. Furthermore, the shortest calculation time is 22.40 (sec), which is the shortest calculation time among the other classifiers.

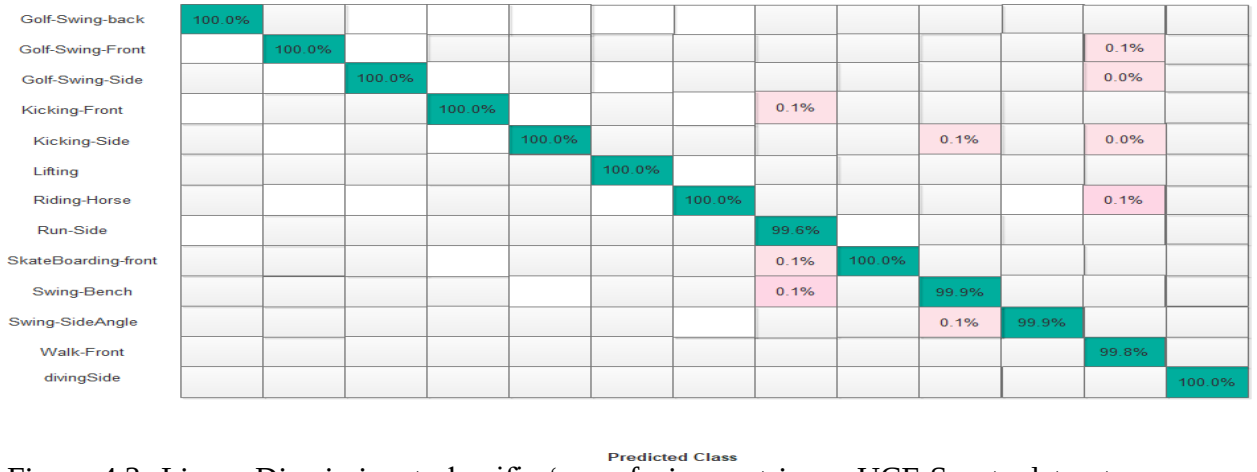


Figure 4.3: Linear Discriminant classifier‘s confusion matrix on UCF Sports dataset using Resnet50 deep features.

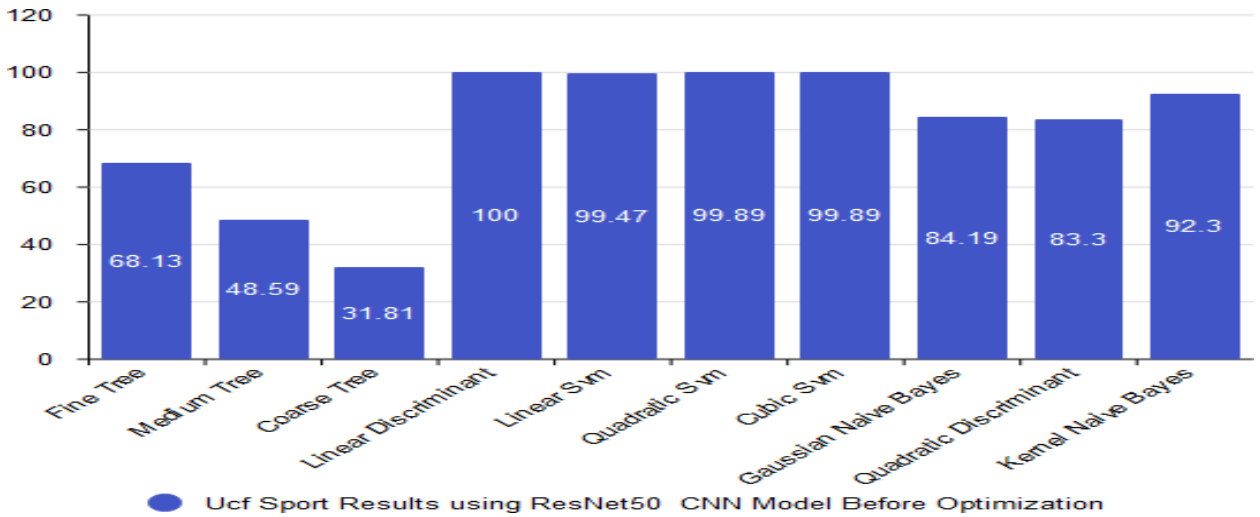


Figure 4.4: Classification computation time on selected classifiers for UCF Sports dataset.

In this experiment, we used the UCFT Youtube dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

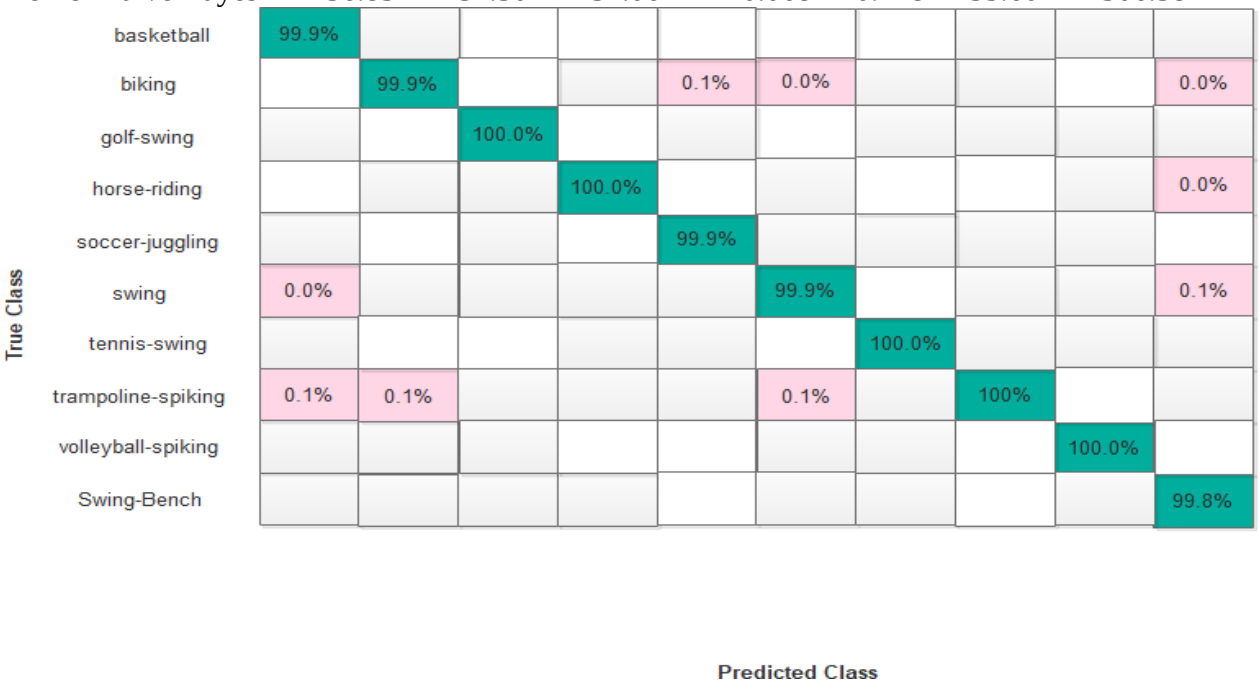
Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 4.1 shows that the Cubic SVM classifier achieved the highest accuracy of 100 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 100, 100, 0.00, and 1.0 being calculated, respectively. The confusion matrix for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Cubic SVM classifier is 583.34 seconds. Furthermore, the shortest calculation time is 56.25 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 3:Feature testing using resnet50 for UCFT Youtube dataset

Classifier	TPR	PPV	F 1-Score	FPR	AUC	ACC%	Time
Fine Tree	73.33	83.78	78.21	0.028	0.937	73.30	133.44
Medium Tree	49.04	71.14	58.06	0.057	0.814	49.10	83.13
Coarse Tree	32.43	32.75	32.59	0.075	0.692	32.40	56.25
Linear Discriminant	99.93	99.94	99.93	0.000	1.000	99.90	83.84
Linear Svm	99.91	99.92	99.91	0.000	1.000	99.90	514.89
Quadratic Svm	100.00	99.99	99.99	0.000	1.000	100.00	551.54
Cubic Svm	100.00	100.00	100.00	0.000	1.000	100.00	583.34
Gaussian Naïve Bayes	82.08	82.13	82.10	0.020	0.933	82.10	113.27
Quadratic Discriminant	99.75	99.76	99.75	0.000	1.000	99.80	99.14
Kernel Naïve Bayes	56.83	57.30	57.06	0.068	0.728	53.60	306.58



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Figure 4.5:Cubic SVM classifier‘s confusion matrix on UCFT Youtube dataset using Resnet50 deep features.

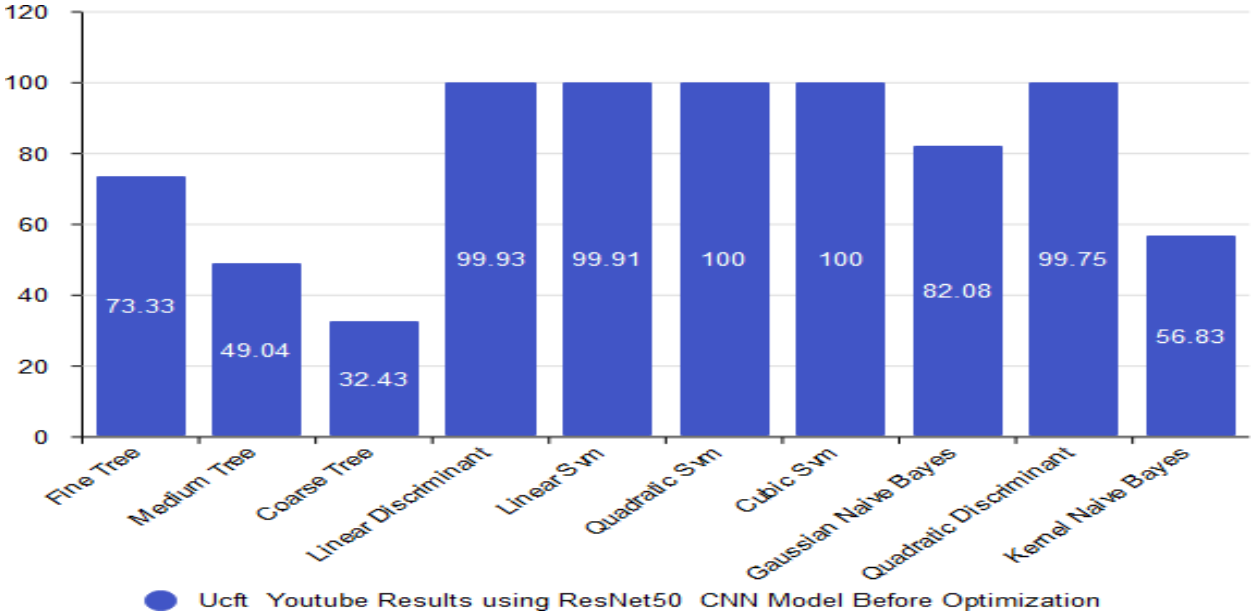


Figure 4.6: Classification computation time on selected classifiers for UCFT Youtube dataset.

In this experiment, we used the Weizman dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Quadratic Discriminant classifier achieved the highest accuracy of 96.10 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 95.44, 98.54, 0.004, and 1.0 being calculated, respectively. The confusion matrix for the Quadratic Discriminant classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Quadratic Discriminant classifier is 289.94 seconds. Furthermore, the shortest calculation time is 8.31 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 4:Feature testing using resnet50 for Weizman dataset

Classifier	TPR	PPV	F 1- Score	FPR	AUC	ACC%	Time
Fine Tree	66.89	67.24	67.06	0.030	0.882	69.20	16.76
Medium Tree	52.15	55.04	53.56	0.127	0.838	55.10	11.44
Coarse Tree	33.85	29.90	31.75	0.069	0.735	38.60	8.31
Linear Discriminant	95.65	95.44	95.54	0.004	0.997	96.10	14.56
Linear Svm	90.88	90.83	90.85	0.008	0.995	91.90	36.10
Quadratic Svm	94.29	94.10	94.19	0.005	0.998	94.90	40.19
Cubic Svm	94.66	94.54	94.60	0.004	0.998	95.20	40.57
Gaussian Naïve Bayes	68.47	68.56	68.51	0.033	0.860	69.30	29.34
Quadratic Discriminant	95.65	95.44	95.54	0.004	0.997	96.10	289.94
Kernel Naïve Bayes	58.31	60.51	59.39	0.003	0.818	56.80	381.54

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

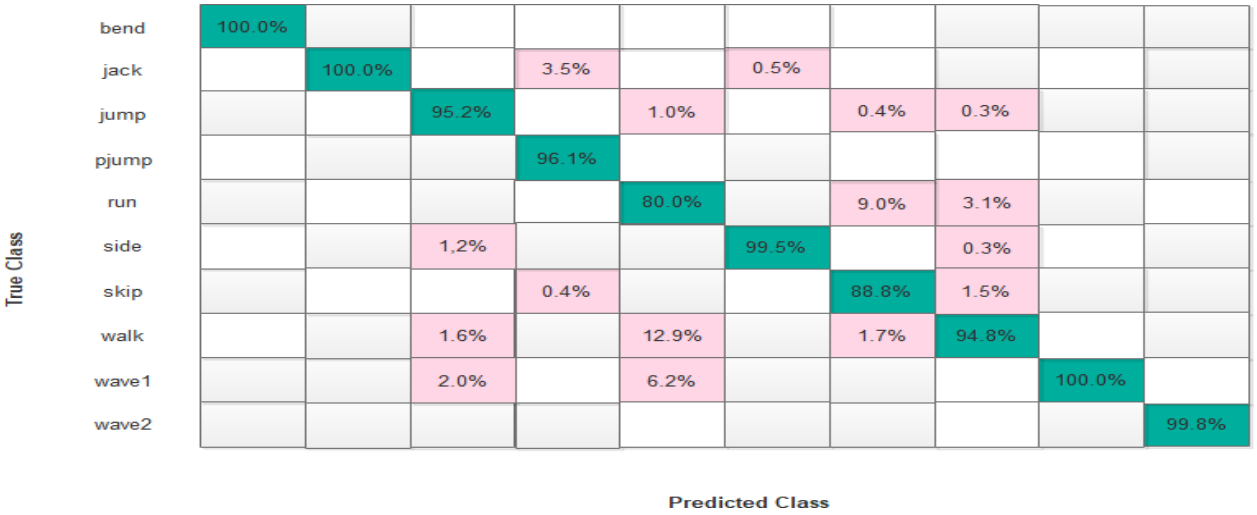


Figure 4.7: Quadratic Discriminant classifier‘s confusion matrix on Weizman dataset using Resnet50 deep features

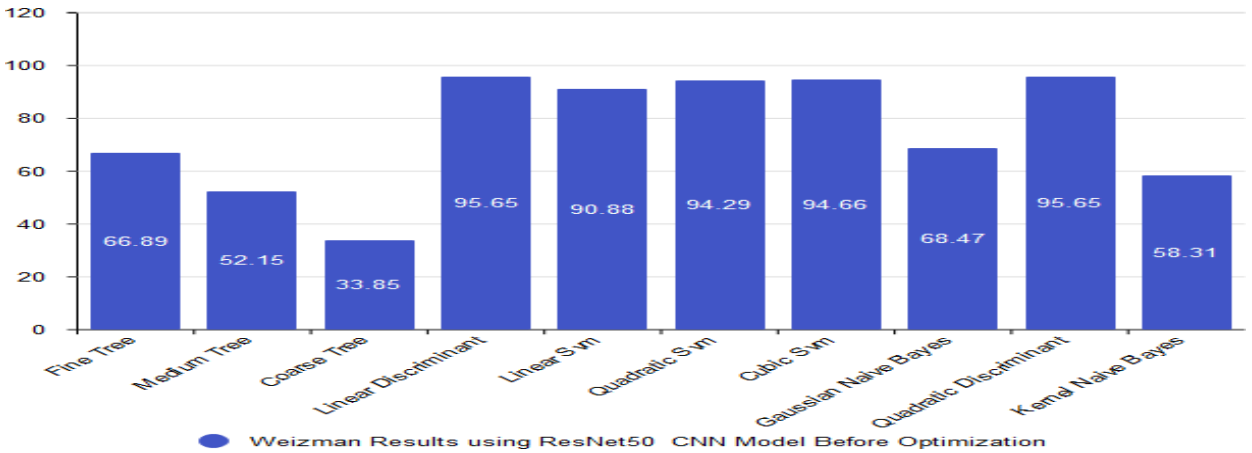


Figure 4.8: Classification computation time on selected classifiers for Weizman dataset.

In this experiment, we used the IXAM dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Cubic SVM classifier achieved the highest accuracy of 85.10 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 84.10, 83.02, 0.014, and 0.987 being calculated, respectively. The confusion matrix for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Cubic SVM classifier is 1029.50 seconds. Furthermore, the shortest calculation time is 94.98 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 5:Feature testing using resnet50 for IXAM dataset Medium

Tree	24.98	32.79	28.36	0.061	0.685	37.30	183.30
Coarse Tree	14.66	8.39	10.67	0.074	0.607	28.80	94.98
Linear Discriminant	80.87	81.35	81.11	0.010	0.961	83.10	98.05
Linear Svm	75.26	80.25	77.67	0.019	0.980	80.40	764.72

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Quadratic Svm	81.88	78.55	80.18	0.014	0.988	85.20	1013.60
Cubic Svm	81.98	84.10	83.02	0.014	0.987	85.10	1029.50
Gaussian Naïve Bayes	34.16	38.10	36.02	0.664	1.328	35.20	201.79
QuadraticDiscriminant	80.87	81.35	81.11	0.016	0.961	83.10	889.16
Kernel Naïve Bayes	37.53	44.59	40.75	0.056	0.699	39.70	7106.80

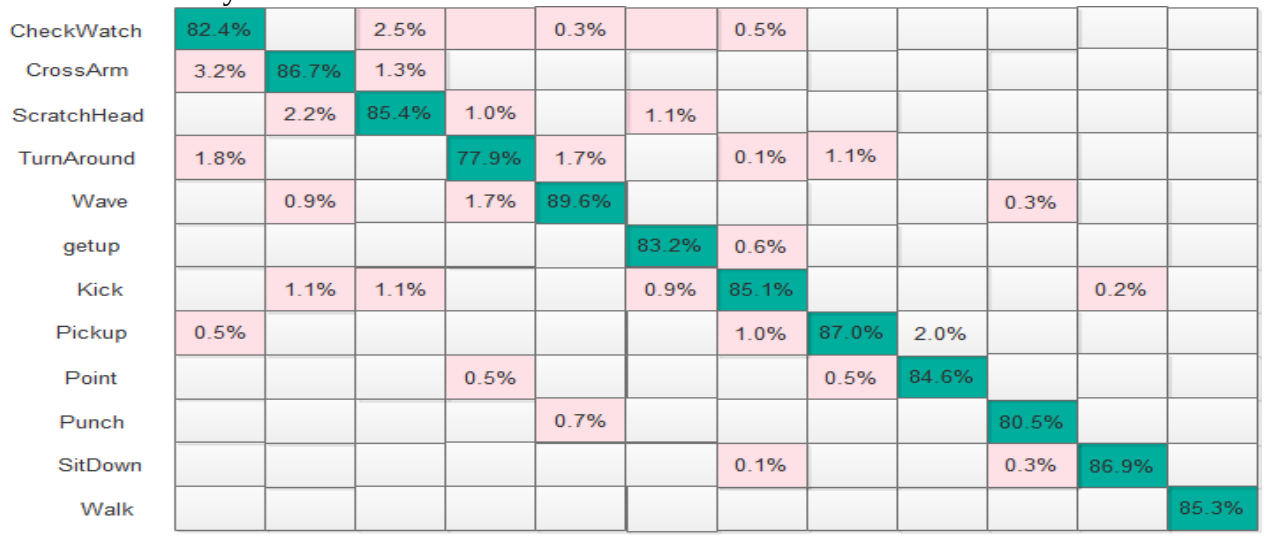


Figure 4.9:Cubic SVM classifier‘s confusion matrix on IXAM dataset using Resnet50 deep features.

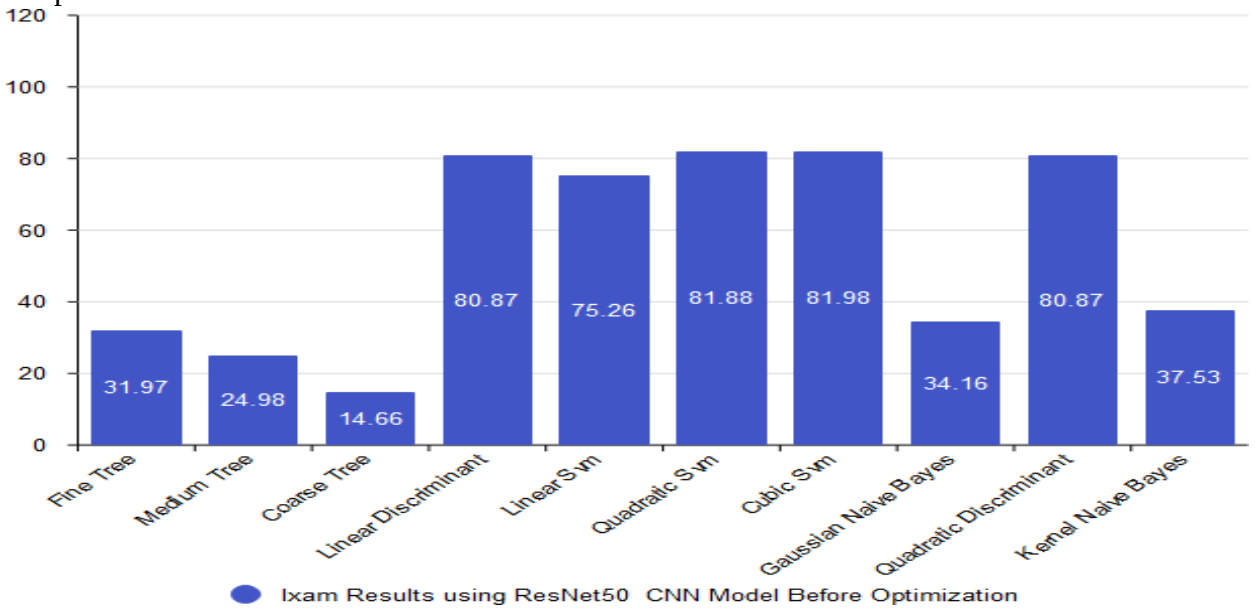


Figure 4.10: Classification computation time on selected classifiers for IXAM dataset.
In this experiment, we used the KTH dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Cubic SVM classifier achieved the highest accuracy of 97.90 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 98.16, 97.97, 0.033, and 1.0 being calculated, respectively. The confusion matrix

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Cubic SVM classifier is 74.96 seconds. Furthermore, the shortest calculation time is 15.55 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 6: Feature testing using resnet50 for KTH dataset After Optimization

Classifier	TPR	PPV	F 1- Score	FPR	AUC	ACC%	Time
Fine Tree	74.06	75.21	74.63	0.052	0.907	74.10	74.02
Medium Tree	63.5	65	64.24	0.075	0.860	63.60	48.00
Coarse Tree	45.4	43.85	44.61	0.106	0.773	46.20	32.00
Linear Discriminant	95.76	95.83	95.79	0.013	0.996	95.60	16.34
Linear Svm	94.38	94.73	94.55	0.011	0.996	94.30	72.84
Quadratic Svm	97.55	97.7	97.62	0.001	1.000	97.50	71.49
Cubic Svm	97.93	98.16	97.97	0.033	1.000	97.90	74.96
Gaussian Naïve Bayes	60.66	62.5	61.57	0.051	0.835	59.80	16.63
Quadratic Discriminant	97.51	97.96	97.73	0.031	0.991	97.60	15.55
Kernel Naïve Bayes	73.38	72.8	73.09	0.056	0.888	72.00	2051.60

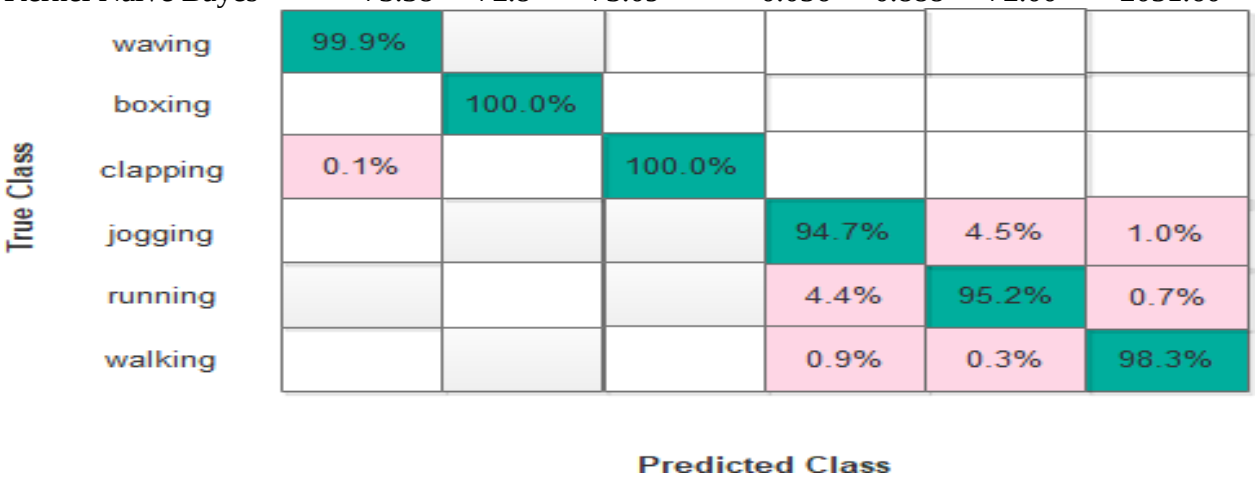
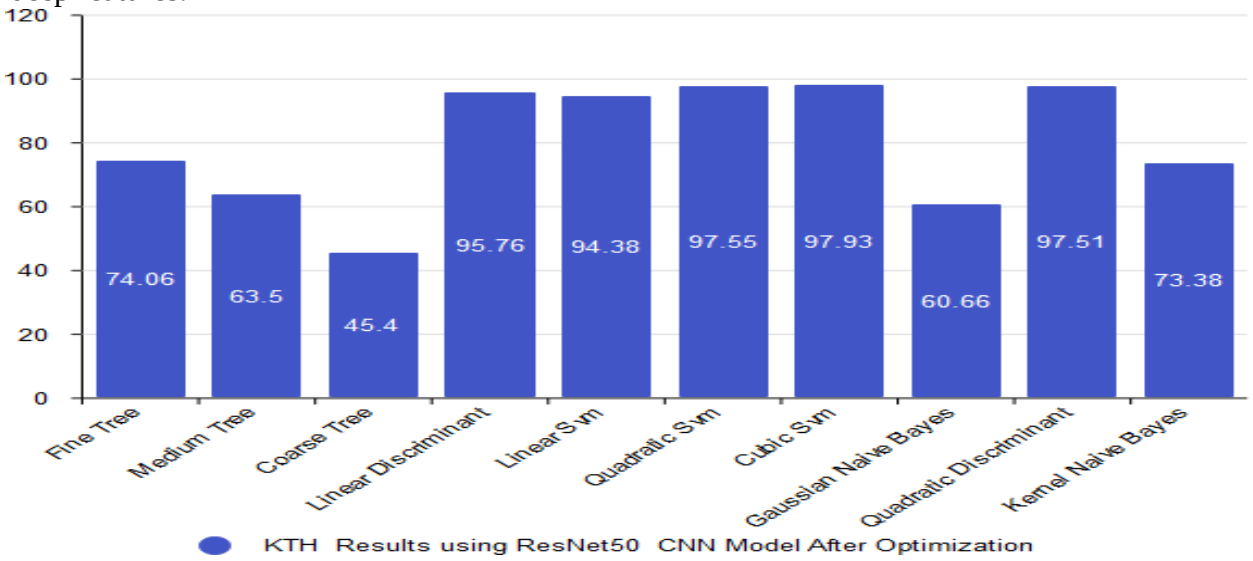


Figure 4.11:Cubic SVM classifier‘s confusion matrix on KTH dataset using Resnet50 deep features.



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Figure 4.12: Classification computation time on selected classifiers for KTH dataset.
In this experiment, we used the UCF Sports dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Linear Discriminant classifier achieved the highest accuracy of 99.90 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 99.81, 99.80, 0.00, and 1.0 being calculated, respectively. The confusion matrix for the Linear Discriminant classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Linear Discriminant classifier is 12.70 seconds. Furthermore, the shortest calculation time is 9.22 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 7: Feature testing using resnet50 for UCF Sports After Optimization

Classifier	TPR	PPV	F1- Score	FPR	AUC	ACC%	Time
Fine Tree	69.77	71.23	70.49	0.031	0.732	81.00	19.12
Medium Tree	49.26	50.13	49.69	0.041	0.871	68.90	13.25
Coarse Tree	30.07	28.91	29.48	0.037	0.835	50.50	9.22
Linear Discriminant	99.79	99.81	99.80	0.000	1.000	99.90	12.70
Linear Svm	99.31	99.72	99.51	0.000	1.000	99.80	70.95
Quadratic Svm	99.89	99.90	99.89	0.000	1.000	99.90	82.61
Cubic Svm	99.73	99.89	99.81	0.000	1.000	99.90	84.23
Gaussian Naïve Bayes	82.71	78.93	80.78	0.018	0.953	56.90	28.7.47
Quadratic Discriminant	84.10	83.21	83.65	0.047	0.671	86.70	124.31
Kernel Naïve Bayes	93.12	92.35	92.73	0.008	0.871	92.10	2131.31

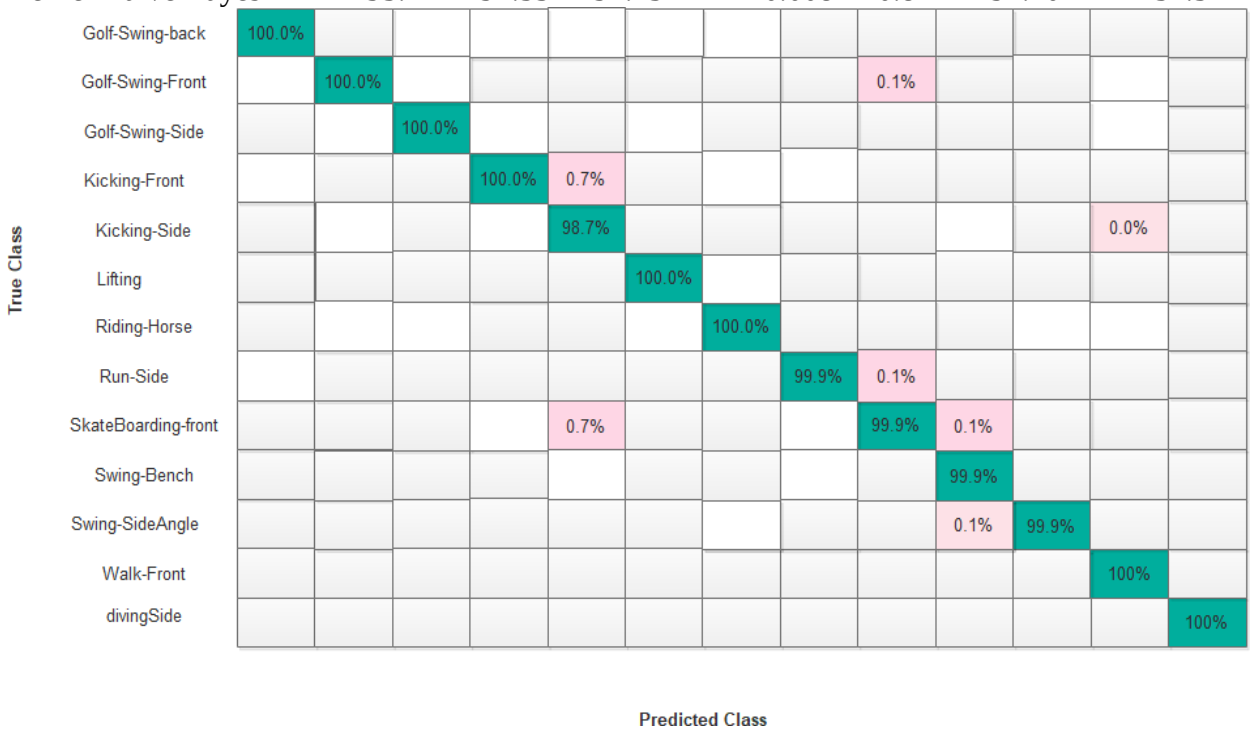


Figure 4.13: Linear Discriminant classifier‘s confusion matrix on UCF sports dataset using Resnet50 deep features.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

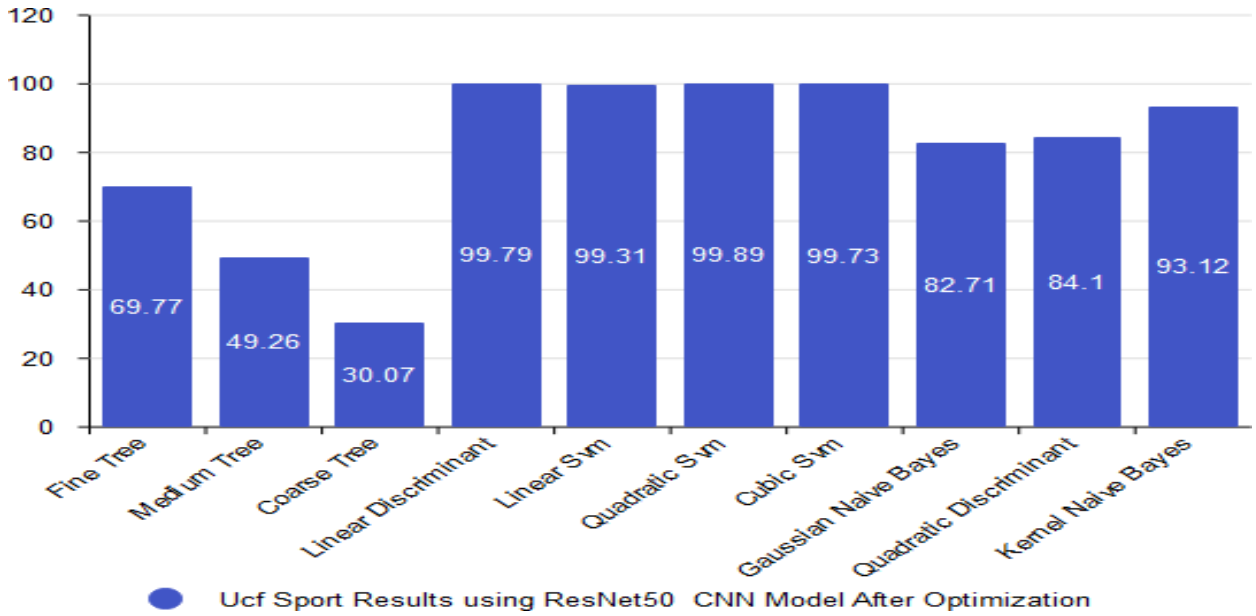


Figure 4.14: Classification computation time on selected classifiers for UCF Sports dataset.

In this experiment, we used the UCFT Youtube dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Cubic SVM classifier achieved the highest accuracy of 100 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 99.99, 99.99, 0, and 1.0 being calculated, respectively. The confusion matrix for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Cubic SVM classifier is 547.32seconds. Furthermore, the shortest calculation time is 50.40 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 8: Feature testing using resnet50 for UCFT Youtube dataset After Optimization

Classifier	TPR	PPV	F 1- Score	FPR	AUC	ACC%	Time
Fine Tree	70.64	75.91	73.18	0.053	0.924	70.60	211.61
Medium Tree	48.95	77.43	59.98	0.055	0.815	48.90	132.78
Coarse Tree	31.67	33.31	32.47	0.076	0.687	31.70	88.28
Linear Discriminant	99.63	99.89	99.76	0.000	1.000	99.60	50.40
Linear Svm	99.85	99.81	99.83	0.000	1.000	99.80	460.80
Quadratic Svm	99.99	99.98	99.98	0.000	1.000	100.00	537.35
Cubic Svm	99.99	99.99	99.99	0.000	1.000	100.00	547.32
Gaussian Naïve Bayes	81.48	81.56	81.52	0.019	0.921	81.50	248.93
Quadratic Discriminant	99.92	99.83	99.87	0.000	1.000	99.90	246.84
Kernel Naïve Bayes	91.40	91.52	91.46	0.051	0.983	91.40	9758.80

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

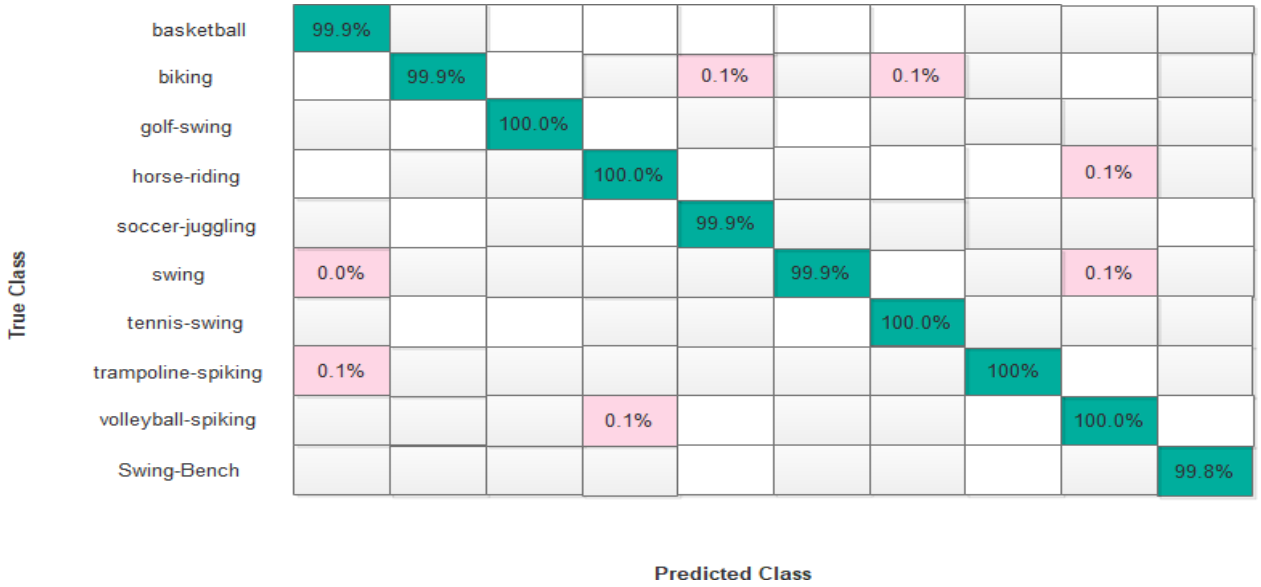


Figure 4.15:Cubic SVM classifier‘s confusion matrix on UCFT Youtube dataset using Resnet50 deep features.

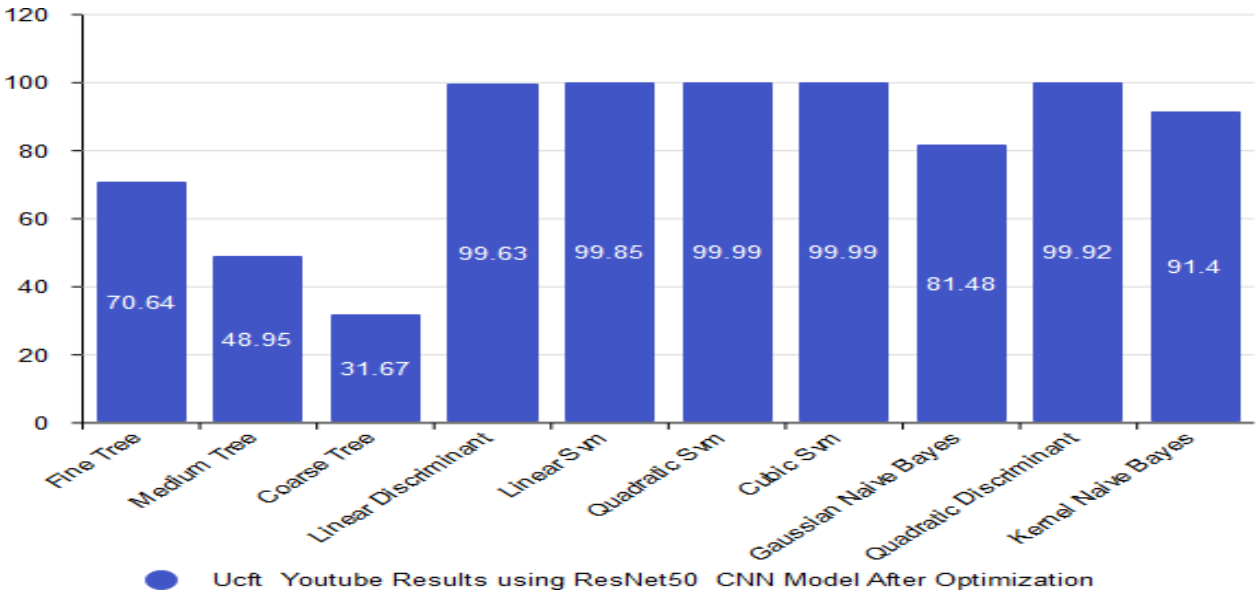


Figure 4.16: Classification computation time on selected classifiers for UCFT youtube dataset.

In this experiment, we used the Weizman dataset dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Cubic SVM classifier achieved the highest accuracy of 95.40 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 94.61, 94.75, 0.005, and 1.0 being calculated, respectively. The confusion matrix for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Cubic SVM classifier is 156.83 seconds. Furthermore, the shortest calculation time is 3.98 (sec), which is the shortest calculation time among the other classifiers.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 4. 9: Feature testing using resnet50 for Weizman dataset After Optimization

Classifier	TPR	PPV	F 1- Score	FPR	AUC	ACC%	Time
Fine Tree	64.84	64.42	64.63	0.028	0.812	66.70	8.05
Medium Tree	53.20	50.87	52.01	0.048	0.831	55.20	5.50
Coarse Tree	33.85	29.80	31.70	0.061	0.721	40.40	4.90
Linear Discriminant	94.56	94.36	94.46	0.005	0.998	95.10	3.98
Linear Svm	90.47	90.41	90.44	0.009	0.997	91.50	22.65
Quadratic Svm	94.20	94.13	94.16	0.006	0.995	94.80	25.15
Cubic Svm	94.90	94.61	94.75	0.005	0.998	95.40	25.67
Gaussian Naïve Bayes	68.77	68.95	68.86	0.037	0.761	69.80	13.72
Quadratic Discriminant	94.20	94.04	94.12	0.007	0.931	94.80	25.15
Kernel Naïve Bayes	73.82	73.20	73.51	0.031	0.723	74.60	452.36

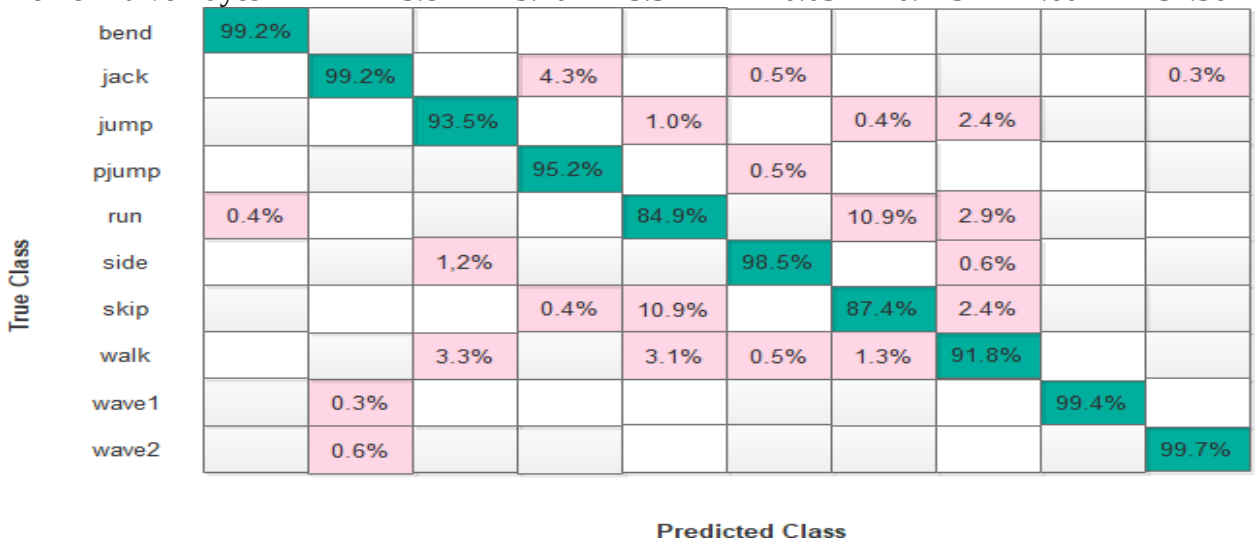


Figure 4.17:Cubic SVM classifier‘s confusion matrix on Weizman dataset using Resnet50 deep features.

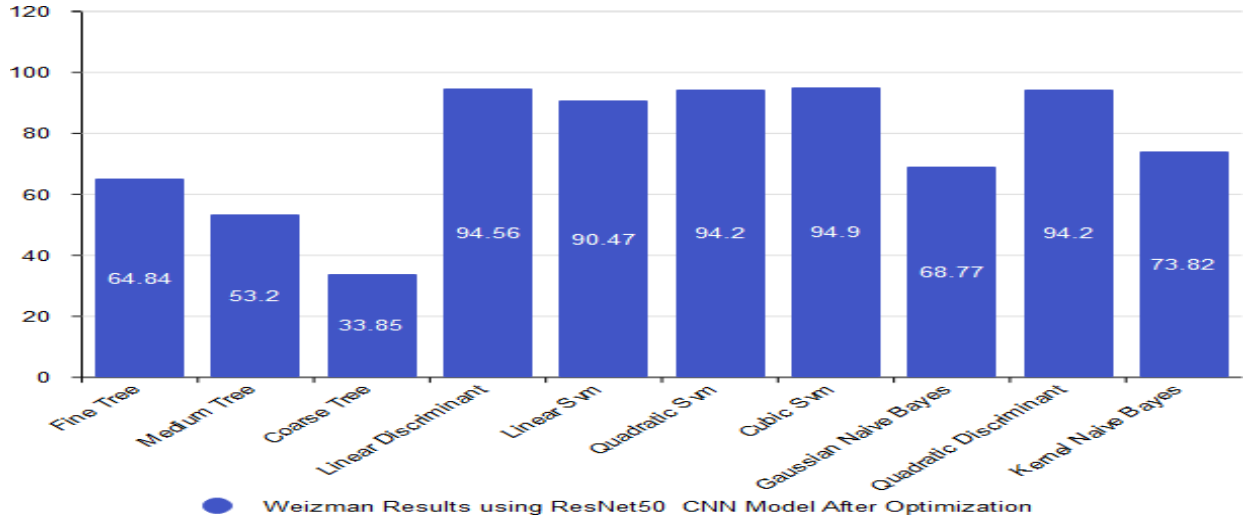


Figure 4.18: Classification computation time on selected classifiers for Weizman

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

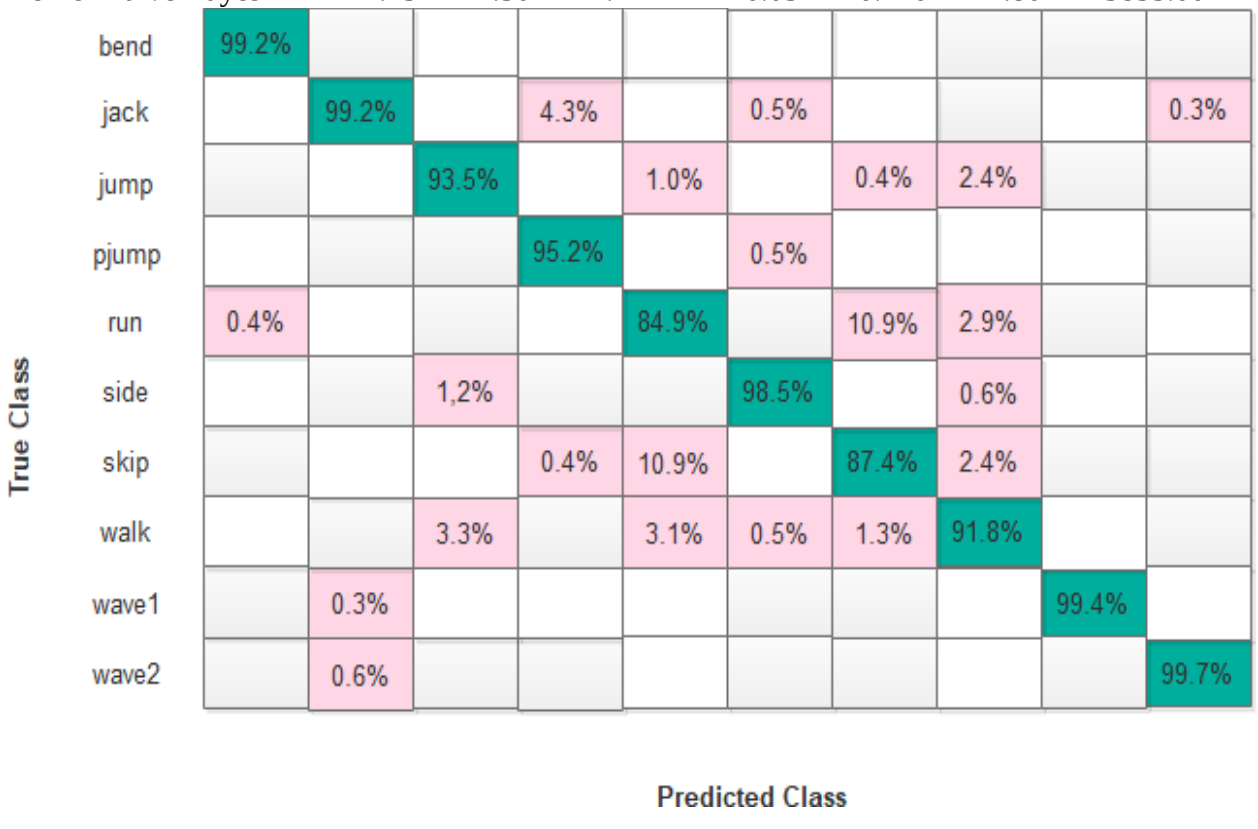
http://amresearchreview.com/index.php/Journal/about

dataset.

In this experiment, we used the IXAM dataset, and features were obtained using the Resnet50 model, which were then fed into the classifiers as a feature vector. Table 4.1 shows that the Cubic SVM classifier achieved the highest accuracy of 92.30 percent, with additional metrics such as precision, F1-Score, FPR, and AUC calculated as 92.07, 92.14, 0.090, and 1.0 being calculated, respectively. The confusion matrix for the Cubic SVM classifier is shown in Figure 4.1, which verifies the accuracy of the classifier. The execution time of the Cubic SVM classifier is 638.32 seconds. Furthermore, the shortest calculation time is 26.53 (sec), which is the shortest calculation time among the other classifiers.

Table 4. 10: Feature testing using resnet50 for IXAM dataset After Optimization

Classifier	TPR	PPV	F 1-Score	FPR	AUC	ACC%	Time
Fine Tree	31.09	34.93	32.90	0.063	0.709	33.40	154.86
Medium Tree	23.74	24.92	24.32	0.064	0.653	27.00	107.85
Coarse Tree	15.08	13.95	14.49	0.076	0.600	19.60	66.44
Linear Discriminant	75.70	75.75	75.72	0.022	0.943	75.90	26.53
Linear Svm	68.52	69.87	69.19	0.025	0.955	70.00	552.75
Quadratic Svm	89.03	89.30	89.16	0.009	0.990	89.40	629.31
Cubic Svm	92.07	92.21	92.14	0.090	0.995	92.30	638.32
Gaussian Naïve Bayes	31.75	37.55	34.41	0.061	0.666	32.20	139.25
Quadratic Discriminant	77.47	90.95	83.67	0.019	0.892	80.80	199.82
Kernel Naïve Bayes	41.75	47.50	44.44	0.052	0.740	41.80	3833.00



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Figure 4.19:Cubic SVM classifier‘s confusion matrix on IXAM dataset using Resnet50 deep features.

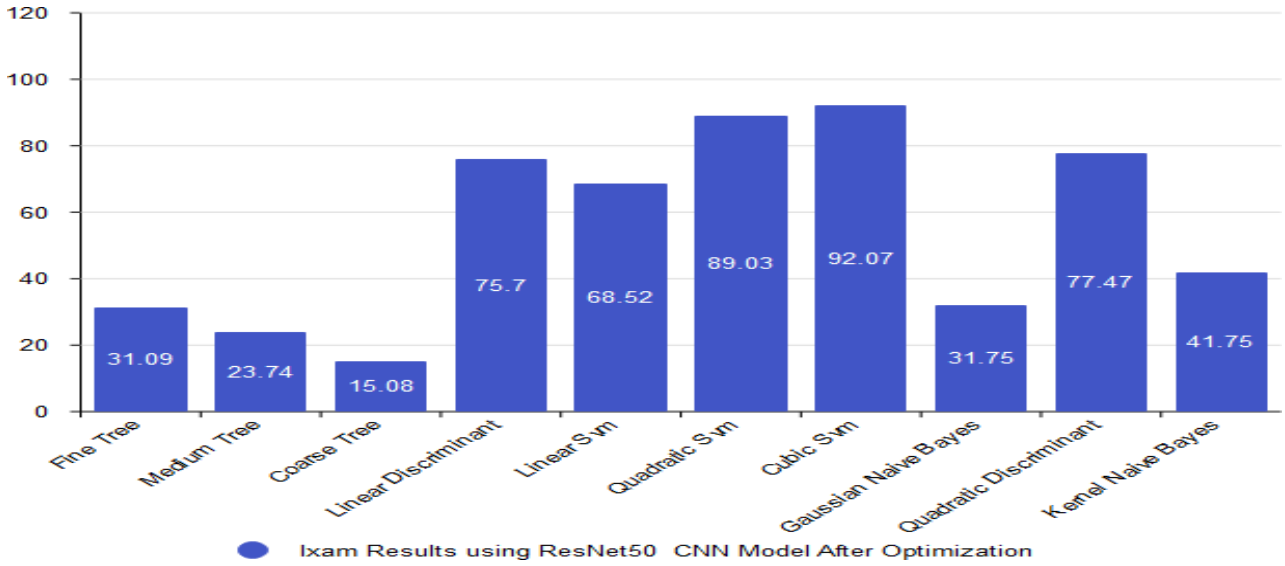


Figure 4.20: Classification computation time on selected classifiers for IXAM dataset.

Conclusion

Human action recognition (HAR) is rapidly gaining popularity in the field of pattern identification due to its unobtrusive features acquired from a distance and the fact that, unlike some other biometric applications like iris and fingerprint detection, it does not require the participation of individuals. It is helpful in a variety of applications, the most significant of which being video surveillance, because of these distinct features. In this work, a new technique is presented for HAR. The suggested methodology's flow architecture includes the following steps: (i) frames initial preprocessing (ii) Inceptionv3 and Resnet50 based deep features extraction (iii) feature fusion (iv) feature selection using IWOA, and (v) classification. Five publicly accessible datasets such as (i) KTH, (ii) UCF Sports, (iii) Weizman, (iv), UCFT Youtube (v) IXMAS dataset are used to validate the HAR method and attained an accuracy of 100%, 99.9%, 99.1%, 99.2% and 100% respectively. The recognition results show that the presented model perform well on these datasets.

IWOA based optimization method are used to select the optimal features for classification by selecting the robust features. This optimization algorithm improves the computational time without decreasing the classification accuracy. In future, we will use some large datasets such as UCF101, Muhavi and HMDB51 to evaluate our presented method. Additionally, to reduce the computational time few other feature selection methods will be used.

References

Ahmad, T., Rafique, J., Muazzam, H., & Rizvi, S. T. H. (2015). Using discrete cosine transform based features for human action recognition. Journal of Image and

Online ISSN 3007-3197
Print ISSN 3007-3189

Graphics, 3(2).

Althloothi, S., Mahoor, M. H., Zhang, X., & Voyles, R. M. (2014). Human activity recognition using multi-features and multiple kernel learning. Pattern Recognition, 47(5), 1800-1812.

Barbu, T. (2014). Pedestrian detection and tracking using temporal differencing and HOG features. Computers & Electrical Engineering, 40(4), 1072-1079.

Bilal, M., Maqsood, M., Yasmin, S., Hasan, N. U., & Rho, S. (2021). A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. The Journal of Supercomputing.

Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding, 117(6), 633–659.

Chenarlogh, V. A., & Razzazi, F. (2019). Multi-stream 3D CNN structure for human action recognition trained by limited data. IET Computer Vision, 13(3), 338-344.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(12), 2247-2253.

Guha, R., Khan, A. H., Singh, P. K., Sarkar, R., & Bhattacharjee, D. (2020). CGA: a new feature selection model for visual human action recognition. Neural Computing and Applications, 33(10), 5267–5286.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Hsieh, C. H., Hung, M. H., & Huang, W. Y. (2016). Recognizing human actions by fusing spatial-temporal HOG and key point histogram. Journal of the Chinese Institute of Engineers, 39(8), 901-909.

Jaouedi, N., Boujnah, N., & Bouhlel, M. S. (2020). A new hybrid deep learning model for human action recognition. Journal of King Saud University - Computer and Information Sciences, 32(4), 447–453.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 221–231.

Liu, J., et al. (2010). Action recognition by multiple features and hyper-sphere multi-class SVM. 20th International Conference on Pattern Recognition.

Mirjalili, S., & Lewis, A. (2016). The Whale Optimization Algorithm. Advances in Engineering Software, 95, 51-67.

Moussa, M. M., et al. (2015). An enhanced method for human action recognition. Journal of Advanced Research, 6(2), 163–169.

Muhammad, K., et al. (2021). Human action recognition using attention based LSTM network with dilated CNN features. Future Generation Computer Systems, 125, 820–830.

Poppe, R. (2010). A survey on vision-based human action recognition. Image and Vision Computing, 28(6), 976–990.

Rodriguez, M. D., Ahmed, J., & Shah, M. (2008). Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. IEEE

Online ISSN 3007-3197
Print ISSN 3007-3189

Conference on Computer Vision and Pattern Recognition.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. Proceedings of the 17th International Conference on Pattern Recognition.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 18(11), 1473–1488.

Weinland, D., Boyer, E., & Ronfard, R. (2006). Action recognition from arbitrary views using 3D exemplars. IEEE International Conference on Computer Vision.

Yuan, Y., Zheng, X., & Lu, X. (2016). A discriminative representation for human action recognition. Pattern Recognition.