Transfer Learning for Classification of Alzheimer's Disease Based on Genome Wide Data

Abbas Saad Alatrany, Wasiq Khan, Abir J. Hussain, Jamila Mustafina, Dhiya Al-Jumeily and for the Alzheimer's Disease Neuroimaging Initiative*

Abstract— Alzheimer's disease (AD) is a type of brain disorder that is regarded as a degenerative disease because the corresponding symptoms aggravate with the time progression. Single nucleotide polymorphisms (SNPs) have been identified as relevant biomarkers for this condition. This study aims to identify SNPs biomarkers associated with the AD in order to perform a reliable classification of AD. In contrast to existing related works, we utilize deep transfer learning with varying experimental analysis for reliable classification of AD. For this purpose, the convolutional neural networks (CNN) are firstly trained over the genome-wide association studies (GWAS) dataset requested from the AD neuroimaging initiative. We then employ the deep transfer learning for further training of our CNN (as base model) over a different AD GWAS dataset, to extract the final set of features. The extracted features are then fed into Support Vector Machine for classification of AD. Detailed experiments are performed using multiple datasets and varying experimental configurations. The statistical outcomes indicate an accuracy of 89% which is a significant improvement when benchmarked with existing related works.

Index Terms— Alzheminer's Disease, GWAS, SNPs, Machine Learning, Transfer Learning, Genome Wide Data

1 Introduction

Alzheimer's disease (AD) is the most common type of dementia with ever increasing prevalence within people over 65 years of age. Despite of significant attempts to study the disease biology and create therapeutic drugs, the cause and course of the disease remain unknown, and there is no treatment available to stop or reverse the disease other than symptomatic treatments [1]. In order to assess efficacy in the development of AD treatments, it is critical to enrol relevant individuals using accurate disease diagnosis techniques. However, clinical diagnosis of AD is based on a physician's assessment of specific neurological and cognitive symptoms, which can be subjective [2].

Generally, the AD can be categorised as Early-onset AD (EOAD) and late-onset AD (LOAD) [3]. An EOAD has been found in about 5% of AD patients with onset ranging from the age of 30s to the mid-60s. Studies have identified the presenilin 1, presenilin 2, and amyloid precursor protein which are the genes involved in EOAD [4]. A LOAD, on the other hand is common one which appears after the age of mid-60s and affects 90–95 percent of total AD patients. Literature indicates Apolipoprotein E (APOE e4) as the frequently confirmed gene being affected in LOAD [5].

The LOAD has been appearing as a complicated condition caused by both hereditary and environmental factors [6]. Because AD has no definitive cure, studying the genes associated in its progression serves as a guide for an early identification of LOAD, close monitoring at risk patients, early treatment and prevention of the disease.

Genome-Wide Association Studies (GWAS) are a frequent study strategy for determining association between common DNA sequence variants and a phonotype. The GWAS studies are large-scale studies that collect genetic diversity in form of SNPs across the human genome. Each of the variations is statistically assessed to find links to a well-defined trait being under investigation [7]. The case-control design is the most prevalent strategy in GWAS, where cases refer to a cohort that has been affected by the disease under study while controls refer to healthy (i.e., normal) subjects. The odds ratio (OR) is the first statistical measure considered in a traditional case-control GWAS, with an OR value greater than 1 indicates the association of an allele is a risk for disease whereas an OR less than 1 indicates the association of an allele as a protective association against the disease [8].

Literature have also reported that the genetic factors play a significant role in AD. In 2013, one of the largest AD

- *A. S. Alatrany is with the Department of Computer Science and Mathematics, Liverpool John Moores University, Liverpool L3 3AF, UK and University of Information Technology and Communications, Baghdad, Iraq. E-mail: a.s.alatrany@2020.ljmu.ac.uk.*
- *W. Khan, A. J. Hussain, and D. Al-Jumeily are is with the Department of Computer Science and Mathematics, Liverpool John Moores University, Liverpool L3 3AF, UK. E-mail: {w.khan, a.hussain, d.aljumeily}@ljmu.ac.uk. Professor Hussain is also with Department of Electrical Engineering, University of Sharjah, Sharjah, UAE.*
- *J. Mustafina is with Kazan Federal University, Kazan, Russia. E-mail: DNMustafina@kpfu.ru.*
- **Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf*

GWASs study reported 19 risk loci related to AD [9]. Recent works also identified additional risk loci (risen to 40) [10-12] which clearly shows the significant contribution of GWAS towards the understanding the genetic components associated with the AD.

Despite the success of classical machine learning (ML) approaches within a wide range of practical applications, it has certain limitations in some real-world dynamics. A typical supervised ML approach require substantial amount of labelled training instances with the same distribution as the test data. However, in many cases, gathering sufficient labelled training data is prohibitively expensive, time-consuming, or even impractical [13]. One of the most commonly used ML approaches to address the aforementioned problem is Transfer Learning (TR), which learns the underlying knowledge required to solve one problem using large amount of data and applies it to subsequent problems with comparatively small datasets. The base network is firstly trained over larger dataset for a certain task which is then used to be fine-tuned over comparatively small dataset in the target domain [13]. Although there are many studies [14-19] using machine learning in the area of GWAS. However, there are some limitations to these studies. In terms of accuracy, results show low predictive performance or a biased performance (i.e., the model's sensitivity is higher than its specificity). While other methods involve only a simple universal test to select relevant features.

In this study, we employ multiple types of transfer learning for the reliable classification of AD using GWAS data. In contrast other existing literature, the proposed study comprises following novelties:

a) To the best of our knowledge, this is the first study to use deep transfer learning to address the data size challenges associated with the GWAS.

b) A comprehensive analysis of multiple types of the transfer learning models has been proposed.

c) Varying configurations of transfer learning applied to GWAS data.

d) A robust feature selection approach to identify the most promising SNPs contributing to the AD classification.

The reminder of this paper is organised as follows. Section 2 represents the related works while Section 3 presents the materials and proposed approach for the AD classification. Results and discussions on statistical outcomes are detailed in Section 4. Finally, Section 5 concludes the findings of proposed study.

2 Rleated Work

Modern ML methodologies employing well-planned AD research can be used to investigate the complexity of LOAD [20]. The major goal has been set to identify and understand various factors that contributes the development of AD. In study [14], the authors investigated three ML claiming as powerful predictive models (i.e., least absolute shrinkage and selection operator (LASSO), step-wise, and genetic algorithm) and suggest that the misclassified data can be used to increase an overall prediction accuracy. The results reveal that adding misclassified sample attributes to the initial model enhanced Area Under the receiver operating characteristic Curve (AUC) by about 5%, reaching to 84%.

To forecast the major depressive illness responses and remissions, different ML models has been proposed using GWAS data [15]; a database comprising 186 patients classed as Major Depressive Disorder (MMD) responders or remitters. LASSO regression was used to extract the most promising variables from a genome-wide association test to discover the possible important variations related to the duloxetine response/remission. Subsequently, support vector machines (SVM) and classification-regression trees were applied to construct the classification models. In relation to duloxetine response, none of the models indicated satisfactory outcomes. The SVM performed comparatively better in terms of remission, producing 52%, 58% and 46% accuracy, sensitivity, and specificity, respectively.

The work presented in study [16] compared different ML models for predicting LOAD from genetic data supplied by the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort in a systematic manner. According to the outcomes, the top performing models generated 72% of AUC towards the classification of LOAD and healthy individuals.

In study [18] utilised DL using convolutional neural networks (CNNs), separated the genome into nonoverlapping fragments and then selected the fragments associated with phenotypes. By significant SNP from the identified fragments build a CNN classification model for AD.

Maj et. al [19] assess the applicability of multiple ML algorithms using omics data from ADNI, which is based on matrices of tissue-specific predicted transcriptome profiles in AD as a case study. Variational autoencoder pre-processing of input data was discovered to be an effective for feature selection prior to the development of classification models using deep learning. The outcomes reported that the Random Forest (RF), Logistic Regression (LR) and SVM were unable to learn to classify cases and controls, because the samples were only assigned to the majority class. The findings also suggest that integration of unsupervised and supervised ML methods can provide complementary knowledge, leading to better performance.

In addition to aforementioned literature, it should also be noted that the AD is largely occur due to genetic causes. As a result, one of the ADNI's main goals is to provide researchers the ability to associate genetics with imaging and clinical data, in order to better understand the disease causes. In this regard, GenADA is a multi-site collaborative effort that aims to create a dataset of 1000 AD patients and 1000 ethnically matched controls in order to analyse the DNA sequence changes in candidate genes with respect to symptoms of AD.

In relation to genetic aspect in AD, study [21] focuses on identifying AD biomarkers using ML techniques. On multiple AD genetic data, the learning algorithms used include Nave Bayes (NB), SVM, LR, and RF. The results show that the overall accuracy of the NB, RF, SVM, and LR learning algorithms is 98%, 97.9%, 95.8%, and 83%, respectively. The findings also indicate that the classifications techniques are beneficial to aid in the early detection of Alzheimer's illness. A similar work is presented in study [22] to predict the AD using SVM model trained over gene-coding protein sequence data. The work used frequency of two successive amino acids to characterise the sequence information. According to the experimental results, the proposed approach for identifying AD indicated an accuracy of 85.7%. The study outcomes also revealed that the sequence information of gene-coding proteins can be used to forecast the AD.

In addition to the ML models employed in above literature, various levels of success have been achieved by kernel functions used in the prediction model to capture nonlinear effects [23, 24]. Nevertheless, kernel-based methods are often sensitive to the underlying aetiology of diseases because their performance is largely determined by pre-selected kernels [25].

2.1 Review of Transfer Learning in Bioinformatics

In transfer learning, a pre-trained model can be used as base model allowing knowledge transmission for given task which is particularly useful to avoid repetitive training [26]. As part of the TL process, knowledge is gained from a dataset (source domain) and transferred to a new dataset (target domain), thereby improving learning in the target domain.

Generally, TL can be categorised into three subcategories that include inductive, transductive, and unsupervised TL. These categories are based on difference in context between the source and the target domains and tasks [27]. The TL has been widely used in various bioinformatic applications [28-30]. Zhao et. al., [31] use TL to propose a polygenic risk score (PRS) method called TL-PRS. The ML model from an ancestry group with large GWAS samples is fine-tuned to fit the target dataset. The model was applied to South Asian and African ancestry individuals from the UK Biobank for seven quantitative and two dichotomous traits. In comparison to the standard PRS method, the TL-PRS method achieved an average relative improvement based on predicted R squared of 25% for South Asian samples and 29% for African samples. Another example of a multi-modal deep learning method in genomics is the DeePathology [32], which uses multi-task and TL to simultaneously infer multiple properties of the biological samples. Using the fine-tuned model, the work reported accurate prediction of tissue and disease types based on the whole transcription profile.

There has been a demonstration of the utility of TL for chromatin accessibility prediction models based on sequences. An analysis of 149 cell types was carried out using the multitask Basset model [33] to predict binary chromatin accessibility profiles. Following this, single-task models of chromatin accessibility were trained using parameters derived from the multitask model. Compared to the models with randomly generated parameters, models with transferred parameters indicated better performance in terms of prediction. However, there are yet several unknowns regarding how many parameters should be shared and which models should be used for which tasks [34].

A recent study [25] presents an explainable ML model for the analysis of high-dimensional genomic data using deep TL. By using the proposed group-wise feature importance score, the study proposes a method for detecting predictive genes harboring both linear and non-linear genetic variants. Using the proposed TL based network architecture, disease risk can also be predicted based on the detected predictive genes. This method was built at the gene level, so it is much easier to interpret the model biologically [25].

In relation to the use of TL in GWAS, a novel statistical method called TL-Multi seeks to improve the polygenic risk prediction across diverse populations, by using summary statistics from GWAS from different ancestries and incorporating the concept of TL [35]. Likewise, Muneeb et. al. [36] proposed prediction of genotype-phenotype with deep learning models through TL while utilising a simulated data.

In contrast to aforementioned literature particularly, the use of TL in GWAS, which either use simulated data or classify AD at gene level, the proposed approach in this study is utilises TL in GWAS analysis on real data. We firstly train a deep CNN model over a GWAS dataset which is then used to extract features from another GWAS AD dataset. The selected features are then fed into a SVM model for the classification of healthy and unhealthy individuals at SNPs level.

3 MATERIALS AND METHODS

The proposed approach exploits TL where multiple datasets are used to train a deep ML model and transfer the learned knowledge efficiently to target domain (to predict the AD class). We conducted detailed experiments to analyse the effectiveness of varying types of TL and to investigate the impact of knowledge transfer from one dataset to another in GWAS analysis. The proposed approach is composed of several components that include quality control, association test, feature selection and classification. A detailed description of each task is provided in the following sections.

3.1 Datasets

The following three datasets comprise the GWAS data sets used in this study:

Dataset A: ADNI GWAS dataset

GWAS dataset requested from the ADNI database (http://adni.loni.usc.edu). The ADNI was founded in 2003 as a public-private cooperation with primary purpose to investigate if magnetic resonance imaging, positron emission tomography, and other biological markers and clinical and cognitive assessments could be used to track MCI and early AD progression.

The ADNI dataset continues collecting participants information that is classified into three classes: Cognitively Normal (CN), Mild Cognitive Impairment (MCI) or Alzheimer's Disease (AD). In our study, only CN and AD classified participants were selected with 216 controls and 183 cases divided into 215 males and 184 females. Participant were genotypes using Illumina Human610-Quad Bead-

Chip Genotyping Platform comprising 620,901 SNPs in total which are stored in a PLINK [37] format file.

Dataset B: AD GWAS Dataset

The second dataset we use in this study is GWAS case-control dataset obtained from [38]. The inclusion criteria for participants is a) who reported themselves to be from European ethnicity, b) according to the National Alzheimer's Coordinating Centre standards, and c) board-certified neuropathologists confirmed late-onset AD in cases and no neuropathology in controls. Furthermore, participants with death age of over 65 years is selected. Plaque and tangle assessment (unique structures that effect cells in the brain which could contribute to the pathophysiology of the disease) conducted on all cases and controls. Samples with a history of stroke, Lewy bodies, or any other neurological disorder were excluded. The final dataset includes 191 males and 173 females partitioned into 176 cases and 188 controls, each with genotyping information for 502,627 SNPs. The DNA of participants were genotyped via Affymetrix GeneChip Human Mapping 500K Array Set. Detailed information regarding the dataset can be found in primary study [38].

Dataset C: AdaptMap goat GWAS dataset

In contrast to above two dataset (containing human records), the third dataset we use in this study is AdaptMap [39] which contains 4653 animals representing 169 populations from 35 countries spread across 6 continents. To genotype the animals, an Illumina GoatSNP50 BeadChip with 53,347 SNPs was used [40]. This dataset has been used to investigate transductive type of transfer learning.

3.2 Quality Control

In the proposed study, individuals and SNPs were subjected to quality control (QC) and filtering procedures in accordance with conventional QC protocols and guidelines as shown in [41] using PLINK software.

For Dataset A, there are 620901 SNPs before genotyping trimming. Based on the Hardy-Weinberg equilibrium (HWE) test, 72490 markers were excluded (with p = 0.1); 61065 markers failed the HWE test in cases, whereas 72490 markers failed the HWE test in controls. The missingness test failed 31368 SNPs (GENO > 0.1). A total of 154598 SNPs failed the frequency test (MAF 0.1). in total, there are 411077 SNPs remained after frequency and genotyping trimming. One individual is removed for low genotyping (MIND > 0.1). After all quality control stages, a total of 398 individuals and 411077 SNPs are left for subsequent analysis.

For Dataset B, the following QC methods were carried out to filter out the genetic markers. SNPs with the genotype missing rate over 5% are eliminated. Likewise, SNPs are filtered for Hardy-Weinberg with a p-value less than 0.001, and the minor allele frequency was less than 0.05. Furthermore, each individual was subjected to QC processes, which consist of a missing genotyping data rate of 0.05, related people, and sex-homozygosity. For subsequent analysis, 356499 SNPs were retained in the samples.

For Dataset C, SNPs and samples are filtered out for missing genotype data (0.1) and the minor allele frequency was less than 0.05. a total of 51117 SNPs and 2765 samples pass filters and QC.

3.3 Association Analysis

Because the resulting information from association analysis varies, it is critical to select the appropriate method for the given context. The association of all SNPs (in Dataset A and Dataset B) within the study with disease status of binary variables (0/1) for case and control patients was assessed using logistic regression under an additive genetic model. The genomic control of logistic regression association test is adjusted to control the population structure. An association test between SNPs and the AD was carried out to decrease the computationally enormous number of genetic variants. The SNPs are sorted in ascending order by p-value, and only the first 5000 SNPs are retrieved for further analysis. Figure 1 depicts the p-values obtained from the association analysis of the AD GWAS dataset using the standard case-control method in a Manhattan plot. The graphic demonstrates that there are two SNPs that have met the GWAS association threshold.

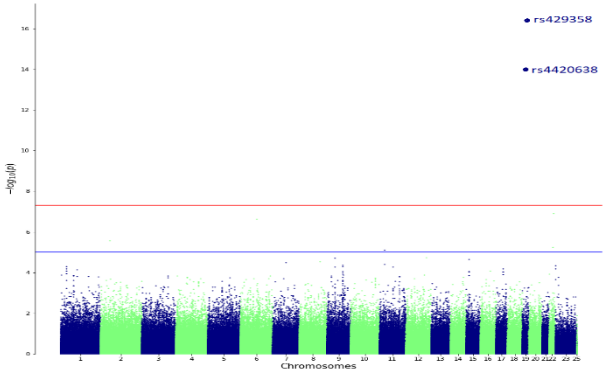

Fig. 1. Manhattan plot of standard case-control shows association of between genotypes and AD.

3.4 Feature selection

The GWAS uses high dimensional data where it is extremely difficult to interpret the data directly, and the majority of the SNPs are irrelevant or uninformative. As a result, identifying the most crucial SNPs is critical. This has three main advantages. Firstly, to simplify the ML model's interpretation. Secondly, it can lower the model's variance and hence overfitting. Finally, reduced number of features can lower the computational cost required to train the ML model. The results of the association analysis are used in this stage to generate a selected features that are significantly associated with the specified phenotype.

The RF is a ML model that has been frequently used for the feature selection [42]. To rank the purity of nodes, RF employs tree-based decision techniques where each decision tree is made up of internal nodes and leaves. The selected features are utilised in the internal node to decide how to partition the data set into two different sets with similar responses. The feature importance is measured as the average of all trees in the forest. The Gini measure, one of the RF methods for measuring feature relevance, used as a feature selector in the current study. Substantial number of SNPs are identified as irrelevant with extremely low significance values. As a result, any SNPs with a Gini value of 0.0009 or higher are included in the feature set for classifications. The significance criterion of 0.0009 was chosen by trial-and-error approach because it can catch the right SNPs that reflect favourable results in the classification task. A total of 60 SNPs were selected by RF as important features which are then used for the classification task for both Dataset A and Dataset B. On the other hand, for Dataset C, only 57 SNPs are used as significant following the Bertolini et. al. [43] analysis of the same dataset. Table 1 shows the top 10 SNPs chosen by RF during the feature selection stage. The SNPs rs429358 and rs4420638 were chosen by RF as two of the top ten features, indicating that the model is effective in identifying the most promising features that are relevant to the disease.

TABLE 1
Characteristics of the top 10 SNPs being selected as important features

SNP	Location	Function	Gene
rs2937774	5:74124992		
rs26642	5:62488562	Intron Variant	IPO11
rs153864	5:62425115	Intron Variant	IPO11
rs7718940	5:86207592		
rs862245	5:82289918	Intron Variant	ATP6AP1L
rs429358	19:44908684	Coding Sequence Variant	APOE
rs4420638	19:44919689	Downstream Transcript Variant	APOC1
rs12374530	5:63761206		
rs37032	5:62388203	Genic Downstream Transcript Variant	KIF2A
rs16890651	5:62333712	Intron Variant	KIF2A

3.5 Convolution Neural Network (CNN) and Transfer Learning

Abstract high-level representation features can be generated using deep learning by combining low-level features, resulting in the finding of data's hidden features. The CNN, one of commonly used deep learning model, can reduce the number of learning parameters by leveraging spatial correlations. As a result, training performance can be improved and data characteristics can be extracted more efficiently [44, 45].

In CNN, the convolution operation extracts the high-level properties such as edges from the input image. The first layer is traditionally in charge of capturing low-level features such as edges, colour, gradient direction, and so on. The architecture adapts to the high-level features as it progresses through the hidden layers, producing a neural network with in-depth understanding of the data [46].

The size of the convolved feature is recued even further in the pooling layer. This is performed to decrease the computational cost required to analyse the data by reducing its dimensionality. Full connection layer is used to learn and merge the non-linear features captured by the preceding layers. The CNN frequently optimises its parameters during the training phase by employing optimisation algorithm (e.g., gradient descent) and modulating the intensity of back-propagation with the learning rate [45].

In order to improve learning in the target domain, TL involves gaining knowledge from a dataset (source domain), and then transferring that knowledge (the pre-trained model) to a new dataset (target domain). The TL is heterogeneous when the source dataset and target dataset come from different domains, with different marginal distributions, predictive distributions, and feature spaces. Homogeneous TL is defined, on the other hand, when the source and target datasets are less different from one another.

In this study, we first time employed both heterogeneous and homogeneous TL to the datasets described earlier. Homogeneous TL was used due to the fact that the human GWAS dataset had the same feature space and domain characteristics from both source and target domains. Source and target datasets of human GWAS used in this study vary in terms of genotyping platforms. Since genotyping platforms tend to generate markers based on a selection strategy and number of markers, the data is influenced by these factors [47]. On the other hand, as the feature space and domain of the animal and human GWAS datasets are different, heterogeneous TL is employed

3.6 Support Vector Machine

Support vector machine (SVM) is a supervised ML algorithm that finds a hyperplane in an N-dimensional space which clearly classifies the input data points. Hyperplane's position and the direction is determined by data points that fall near the hyperplane. Using these support vectors, the classifier's margin is maximised. The position of the hyperplane changes if the support vectors are removed. These are the points that assist in developing the SVM. The bigger the margin of the hyperplane, the more confident algorithm is in classifying new data points [48].

3.7 Experiment Design

The outline of the proposed model is depicted in Figure 2. The GWAS data is pre-processed and filtered to contain only high-quality samples and markers by employing appropriate quality control processes on all datasets. A logistic regression-based association test is performed to identify the SNPs that are strongly associated to the disease. Furthermore, RF algorithm is used for the selection of important features and to reduce the dimensionality, making the number of features appropriate to the number of available observations.

As a result of trial-and-error testing, we chose convolution layers between 2-4 because using excessive layers may overfit the model, while very few layers may limit its capabilities [49]. According to best practices in the literature and similar related work [18, 50, 51], we selected the number of convolutional layers in this study, as well as the other hyperparameters for the deep learning models [52-54]. Where as for SVM and RF classifiers, a grid search is conducted on user defined hyper-parameters values. The deep learning models' structures are demonstrated in Table 2. Customization of Transfer Learning:

- Prediction: use the pre-trained model to immediately classify new observations
- Fine-tuning: unfreeze the classifier, or a portion of it, and retrain it on new dataset.
- Feature extraction: The output of the layer preceding the final layer is fed into a new model as input. The goal is to pre-process the inputs and extract essential features using the pre-trained model, or a subset of it.

Following the completion of the necessary data processing and filtering, multiple experiments are conducted to examine the effectiveness of TL in GWAS domain:

Experiment 1 (EXP1): Implementation of transductive TL, to train the model using source and target dataset from similar domain and a similar task in both source and target models. Therefore, a CNN is trained on Dataset A as a Base CNN where Dataset A is partitioned as 80 percent for training and 20 percent for testing (pre-trained). A number of architectures are built, and the one with the best performance is kept as the base CNN model; the architecture of the base CNN is displayed in (Table 2.A). Then the base CNN is used in three approaches for prediction, fine-tuning, and as a feature extractor for Dataset B.

Experiment 2 (EXP2): Implementation of inductive TL to train the model on source and target datasets from similar domain but different tasks (in our case, we train the model on GWAS data Dataset C of animal to classify goat into 11 subcontinental breeds; the architecture of the base CNN is displayed in Table 2.B) then use the TL over the pretrained model to classify the individuals in Dataset B.

Experiment 3 (EXP3): Implementation of inductive TL to use the pretrained model from EXP2 to classify individuals in Dataset A.

Experiment 4 (EXP4): Implementation of inductive TL to use the pretrained model from EXP2 to classify individuals in an aggregated dataset comprising both Dataset A and Dataset B.

In all of the above experiments, ML algorithms are built using the Scikit-learn Python library [55]. The PyPlink library [56] is used to read the genotype data in Python. Deep learning models are built utilizing Keras and TensorFlow as backend [57].

TABLE 2
ARCHITECTURES OF THE PROPOSED CNNS; (A) FOR EXP1 AND (B) FOR EXP2, 3 AND 4

CNN Model A		**CNN Model B**	
Layer Type	Description	Layer Type	Description
Conv1D	F = 16, K = (5,), ReLu	**Conv1D**	F = 16, K = (5,), ReLu
Conv1D	F = 16, K = (3,), ReLu	**Conv1D**	F = 32, K = (3,), ReLu
Pool1D	Max Pooling (2,)	**Pool1D**	Max Pooling (2,)
Dropout	10%	**Dropout**	10%
Reshape	Flatten	**Conv1D**	F = 32, K = (3,), ReLu
Dense	F = 64, Sigmoid	**Pool1D**	Max Pooling (2,)
Dropout	10%	**Dropout**	10%
Dense	F=2, softmax	**Conv1D**	F = 32, K = (3,), ReLu
		Pool1D	Max Pooling (2,)
		Dropout	10%
		Reshape	Flatten
		Dense	F = 64, Sigmoid
		Dropout	10%
		Dense	F=2, softmax

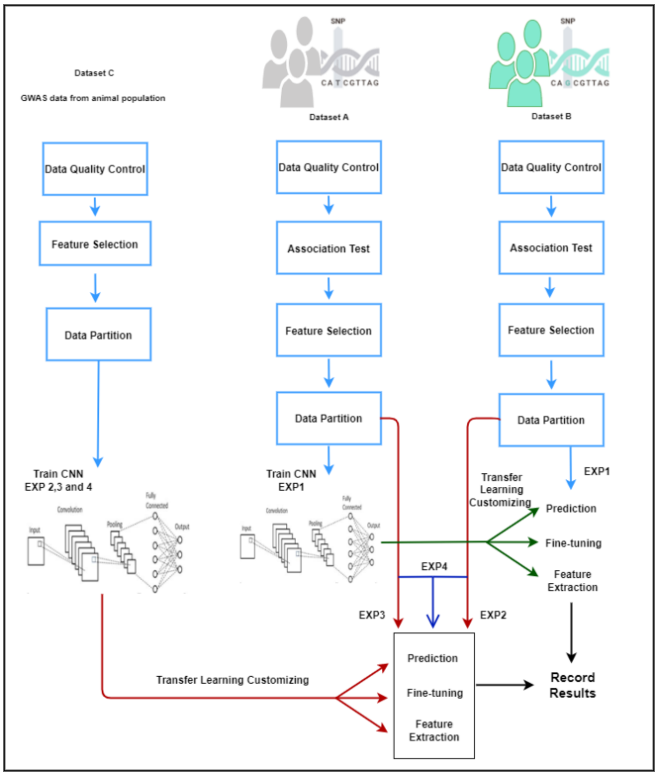

Fig. 2. The proposed Transfer Learning Framework. On left side, quality control and feature selection are conducted on Dataset C, then a CNN is trained on animal data as a base model to be transferred to both Dataset A and Dataset B for EXP 2,3 and 4. In the middle, a CNN model is trained on human data as a base model to be transfer to Dataset B for EXP 1.

4 RESULTS AND DISCUSSIONS

4.1 Evaluation Criteria

In this study, we use GWAS data to train a deep TL model to distinguish the healthy and LOAD-infected subjects. To assess the performance of proposed approach, we use AUC which is one of the commonly used ML evaluation metric [58-61]. Along with the AUC, we employ standard evaluation metrics [62, 63] including accuracy, precision, recall, and F1 score.

4.2 Transductive Transfer Learning Based AD classification (EXP1)

A CNN model is trained and tested over Dataset A in EXP1. The pre-trained model was saved for TL so that it could be reused in the target domain, Dataset B. The pre-trained model firstly used after training only the fully connected layers, to predict the samples in Dataset B. Secondly, we then unfreeze the frozen pre-trained model's layers, then trained the transfer model on 80% of the observations in Dataset B and tested on the remaining observations in Dataset B. Finally, the fine-tined model is used as a feature extractor and serves as an input to ML classifiers (i.e., SVM and RF in this case).

The results attained during this experiment are listed in Table 3 which shows that the highest accuracy (89.04%) and F1 score (88.57%) are achieved by customizing the pre-trained model as a feature extractor and fed into an SVM with rbf kernel. Whereas, utilizing the pre-trained model for the prediction task did not generalize well on target dataset and showed a significant decrease in accuracy to 39%. Although the drop in accuracy, but model achieved high score in terms of recall, in comparison to other models in EXP1. This suggest that accuracy metric is not enough to examine the true performance and any biasedness towards a specific class in a model.

It can be noted that the change in kernel type also influence the model's performance, an improvement of around 2% in terms of both accuracy and f1-score when utilizing rbf kernel compared to liner kernel. Likewise, more balanced performance is achieved using FE+SVM with rbf kernel in terms of precision and recall which is not the case otherwise.

TABLE 3
Results of EXP1 Transductive Transfer Learning (Transfer from Dataset A to Dataset B)

Model Use	Accuracy	Precision	Recall	F1-score
Prediction	0.3972	0.4354	0.75	0.551
Fine-tuning	0.7671	0.8064	0.6944	0.7462
FE+RF	0.8904	0.966	0.8055	0.8787
FE+SVM with linear Kernel	0.8767	0.9354	0.8055	0.8656
FE+SVM with rbf Kernel	0.8904	0.9117	0.8611	0.8857

4.3 Inductive Transfer learning Based AD Classification (EXP 2, 3 and 4)

In EXP2, the source dataset used is GWAS data of animals to train a CNN model to classify the goat into 11 subcontinental breeds. The pretrained model, as in EXP1, adapted to classify the samples of target dataset (Dataset B) by a) only changing and training the top layer, b) fine-tune the model to make them relevant for the target task, c) as a feature extractor. The detailed statistical outcomes of this experiment are shown in Table 4. Similar to EXP1 results, the pre-trained model generalized well when fine-tuned and used as feature extractor followed by an SVM with rbf kernel. However, the model shows maximum accuracy of 60.27% when used directly (without fine-tuning the pre-trained model's layers) to predict the class in the target dataset. This is a significant drop in model's performance which clearly indicates the usefulness of fine-tuning of TL for the task of AD classification. Even though a high accuracy of 84% achieved after fine-tuning and utilizing the pre-trained model to classify samples in Dataset B, there is clearly a biased performance in terms of precision (93%) and recall (75%) metrics which shows biasedness towards one class. In construct, balanced performance of 87% and 80% for precision and recall achieved when customizing the pre-trained model as feature extractor followed by an SVM.

TABLE 4
Results of EXP2 (Transfer from Dataset C to Dataset B)

Model Use	Accuracy	Precision	Recall	F1-score
Prediction	0.6027	0.6060	0.5555	0.5797
Fine-tuning	0.8493	0.9310	0.75	0.8307
FE+RF	0.8082	0.8437	0.75	0.7941
FE+SVM with linear Kernel	0.7671	0.7878	0.7222	0.7536
FE+SVM with rbf Kernel	0.8493	0.8787	0.8055	0.8405

In Experiment 3, the same pre-trained model from Exp 2, is used for the TL over Dataset A. The main objective is to investigate if the pre-trained model is able to well generalize for different datasets. Following the same TL strategies, Table 5 lists the statistical results from EXP3. Unlike the outcomes from EXP2, the pre-trained model did not perform well in general, however, indicates better recall scores than precision. After fine-tune the model to make it more relevant to Dataset A, we achieved 67.5% and 59.37% accuracy and f1-score, respectively. These statistical outcomes clearly indicate that employing the pre-trained model as a feature extractor could not help in improving the model performance in this experiment.

Similar to outcomes from EXP1 and EXP2, the rbf kernel outperforms linear kernel which may be due to the non-linear nature of the dataset.

TABLE 5
Results of EXP3 (Transfer from Dataset C to Dataset A)

Model Use	Accuracy	Precision	Recall	F1-score
Prediction	0.5875	0.4193	0.4642	0.4406
Fine-tuning	0.6750	0.5277	0.6785	0.5937
FE+RF	0.6375	0.4827	0.5	0.4912
FE+SVM with linear Kernel	0.625	0.4705	0.5714	0.5161
FE+SVM with rbf Kernel	0.65	0.5	0.5357	0.5172

In EXP2 and EXP3, the pre-trained model is reused in target domains of Dataset A and Dataset B individually, to examine the generalization of pre-trained model on both datasets. Furthermore, the pre-trained model utilized in both EXP2 and EXP3, is employed as base model to be fine-tuned over aggregated dataset of A and B. the main intention is to investigate how the pre-trained model will behave in varying settings. Table 6 demonstrate the results achieved through this experiment (EXP4). The statistical outcomes show that the accuracy dropped to 58% when customizing the pre-trained model over the aggregated dataset. This might be because of the effect of Dataset A, as the model in EPX3 did not perform very well. Similar to EXP2, the model achieved highest performance of 69.28% and 64.66% of accuracy and f1-score, respectively, when fine-tuned over the aggregated dataset and used as a feature extractor followed by an SVM.

TABLE 6
Results of EXP3 (Transfer from Dataset C to aggregated dataset of Dataset A and dataset B)

Model Use	Accuracy	Precision	Recall	F1-score
Prediction	0.5882	0.5744	0.3857	0.4615
Fine-tuning	0.6601	0.6551	0.5428	0.5937
FE+RF	0.6405	0.6315	0.5142	0.5669
FE+SVM with linear Kernel	0.6666	0.6727	0.5285	0.5920
FE+SVM with rbf Kernel	0.6928	0.6825	0.6142	0.6466

In Fig. 3, we present a comparison of receiver operating characteristic curves (ROC) for each model within each of the four experiments. The best performance was achieved by using transductive transfer learning utilized in EXP1 (Fig. 3a). The second-best results were obtained with inductive transfer learning (i.e. transfer from Dataset C to Dataset B) as demonstrated in Fig. 3b. However, this high performance did not hold when the pre-trained model was transferred to Dataset A (Fig. 3c). Figure 3 also shows that using pre-trained models as feature extractors provides better results than other TL approaches. It was fine-tuning the pre-trained model that resulted in a better AUC in experiment 3.

4.4 Comparison with Related Work

Table 7 presents performance comparison between the proposed TL based AD classification approach and related works from the literature. It can be noticed that our approach outperforms the existing methods in terms of almost all performance metrics with an increase of 5% of accuracy and AUC, and 8% increase in f1-sore. In addition, it is very important to note that the proposed approach uses only 60 features as input to ML model as compared to state of the art [14] which uses over 500 features. This results the proposed model less noisy, light weight, and efficient model. Furthermore, identification of fewer most contributing feature to AD might be useful to set a baseline for further analysis and future research direction. Our proposed model not only show high accuracies, but also shows well balanced performance in terms all metrics. In contrast, gradient boosted decision tress [64] showed an increase of 11% in terms of AUC comparing to other metrics.

4.5 Discussions

The genetics of phenotypes such as AD is of complex nature. Multiple genetic markers play a role in the emergence of complicated human disease. Despite the fact that GWAS were successful in identifying SNPs associated with complex features, this strategy lacks the identification of variants with low influence that might play a significant role when combined with other variants [65]. Additionally, traditional GWAS have only discovered SNPs that can only account for 33% of the estimated 79% [66] of genetic risk related with AD.

Although this value is insufficient for a reliable clinical prediction, ML algorithms have been shown to be more effective in discovering candidate SNPs and predicting complicated genetic diseases [67-69]. In the last decade, the application of ML-based techniques for genetic-based precision medicine has expanded and is expected to continue [70].

The results shown in Table 3 that the TL can be an effective tool for GWAS data classification. This is owing to the fact that the deep learning models must be trained on a large amount of data. Because the high dimensionality of GWAS data makes the training of deep learning models more challenging, transfer learning from one dataset to an other can help to resolve this issue. However, carful selection of source dataset for the pre-trained model plays a major role towards the model performance when the model is transferred to another dataset.

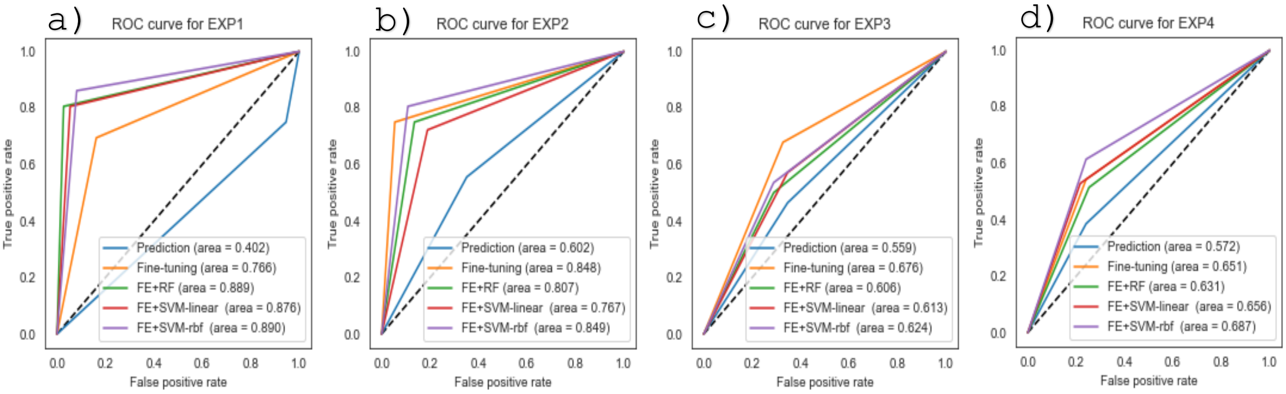

Fig 3. ROC Curves: a) ROC for EXP1, b) ROC for EXP2, c) ROC for EXP3 and d) ROC for EXP4.

TABLE 7
Comparison of related work in the literature

Study	ML model	Dataset	Feature No.	Accuracy	Precision	Recall	F1 score	AUC
[64]	gradient boosted decision trees	UK-BioBank	145	80%	80%	80%	80%	91%
[18]	1D CNN	ADNI	4000	75%	-	-	-	81%
[16]	Ensemble of several ML algorithms	ADNI	2500	~70%	-	70%	-	72%
[14]	LASSO	NIA-LOAD	501	84%	-	82%	-	84%
Propsed Model	Transfer Learning + SVM	ADNI	60	89%	91%	86%	88%	89%

As shown in EXP1 (Table 3), the TL in a similar domain and task, in source and target datasets settings (from Dataset A to Dataset B) indicated the best performance. Although there was a slight difference in the type of population in two data sets (Dataset A from European population, whereas Dataset B contains non-Hispanic participants), still the pre-trained model generalize very well on the target dataset. As majority of the GWAS data comprises European participants [71], this will pave the road for research of minor population. Particularly, where limited GWAS data exists, proposed approach might be effective to use. When using a GWAS data from animal population in light of the similarities in biological function among species [72], TL the pre-trained model was effective in classifying participants in Dataset B, but did not perform well on Dataset A (Table 5). This may be because of the selection of genotyping platform, as the data is known to be influenced by the selection strategy and number of markers generated by genotyping platforms [47]. This requires more investigation to clarify why the model was able to perform well on one dataset but not on another similar one. Results shows that the classification accuracy was reduced when the pre-trained model used for aggregated dataset. This suggests that the pre-trained model may have failed to learn GWAS-specific features, and instead, relying on dataset-specific features.

Three TL customizations are used for the classification of AD. Except one experiment (EXP3), the other three experiments demonstrated the pre-trained model utilized as feature extractor followed by a ML model outperform other customizations. In only one experiment 3, fine-tuning the pre-trained model had better accuracy than other two strategies. This implies that repurpose the previously learned feature maps (from source domain) for the target dataset can help in achieving better performance with TL.

It is also noticed that the RF found to be capable of selecting SNPs that have previously been linked to AD. As a result, SNP selection based on RF could be a useful tool for identifying clinically important risk factors. The current findings backed with previous SNP findings showing the APOE 4 gene is the primary risk factor for LOAD [73].

For highly accurate clinical diagnostic, the genetic component alone forms a barrier. Complementing the genetic-based approaches with imaging or clinical data could be one of the possible answers to this challenge. The genetic study might be used to identify subjects who are at a higher risk of acquiring AD and therefore, such subjects can be tracked with imaging technology on regular basis to detect the disease's onset as soon as feasible [16].

Alongside the proposed study's contributions, small sample size of dataset limits this study; we expecting that increasing the sample size will increase the forecasting performance of the deep TL models. As a result, we are predicting that these models have a lot of potential for diagnosing LOAD and other complex diseases.

5 Conclusion

The outcomes of utilising TL followed by the support vector machine, to estimate the risk of acquiring Late-Onset Alzheimer's Disease entirely from genetic variation data, were presented in this research work. Comparing the classification performance of machine learning models is a critical component of the validation process for the proposed model. The feature selection methodology utilized to decrease the large number of SNPs has the potential to lead to the discovery of new disease-related genetic markers. We expect that the proposed methodology could be a strong tool for classification of AD, based on the preliminary results. This article also shows that transfer learning is an effective method for analyzing and leveraging a large number of genetic markers that might be utilized to a variety of complicated disorders like Alzheimer's. Transductive transfer learning is utilized as a feature extractor, which is found to result in the highest classification performance in this study when compared with other types and customizations of transfer learning.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis

Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] L.-K. Huang, S.-P. Chao, and C.-J. Hu, "Clinical trials of new drugs for Alzheimer disease," *Journal of biomedical science,* vol. 27, no. 1, pp. 1-13, 2020.

[2] Z. S. Khachaturian, "Diagnosis of Alzheimer's disease," *Archives of neurology,* vol. 42, no. 11, pp. 1097-1105, 1985.

[3] P. G. Ridge, M. T. W. Ebbert, and J. S. K. Kauwe, "Genetics of Alzheimer's Disease," *BioMed Research International,* vol. 2013, p. 254954, 2013/07/25 2013, doi: 10.1155/2013/254954.

[4] R. Sims, M. Hill, and J. Williams, "The multiplex model of the genetics of Alzheimer's disease," *Nature neuroscience,* vol. 23, no. 3, pp. 311-322, 2020.

[5] Q. Zhang *et al.*, "Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture," *Nature Communications,* vol. 11, no. 1, p. 4799, 2020/09/23 2020, doi: 10.1038/s41467-020-18534-1.

[6] G. Rabinovici, "Late-onset Alzheimer disease. CONTINUUM: Lifelong Learning in Neurology, 25 (1), 14–33," ed, 2019.

[7] W. S. Bush, "Genome-Wide Association Studies," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach Eds. Oxford: Academic Press, 2019, pp. 235-241.

[8] J. Yang *et al.*, "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits," *Nature genetics,* vol. 44, no. 4, pp. 369-375, 2012.

[9] J.-C. Lambert *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Nature genetics,* vol. 45, no. 12, pp. 1452-1458, 2013.

[10] I. E. Jansen *et al.*, "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk," *Nature genetics,* vol. 51, no. 3, pp. 404-413, 2019.

[11] R. E. Marioni *et al.*, "GWAS on family history of Alzheimer's disease," *Translational psychiatry,* vol. 8, no. 1, pp. 1-7, 2018.

[12] B. W. Kunkle *et al.*, "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing," *Nature genetics,* vol. 51, no. 3, pp. 414-430, 2019.

[13] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE,* vol. 109, no. 1, pp. 43-76, 2020.

[14] B.-L. Romero-Rosales, J.-G. Tamez-Pena, H. Nicolini, M.-G. Moreno-Treviño, and V. Trevino, "Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling," *PloS one,* vol. 15, no. 4, p. e0232103, 2020.

[15] M. Maciukiewicz *et al.*, "GWAS-based machine learning approach to predict duloxetine response in major depressive disorder," (in eng), *J Psychiatr Res,* vol. 99, pp. 62-68, Apr 2018, doi: 10.1016/j.jpsychires.2017.12.009.

[16] J. De Velasco Oriol, E. E. Vallejo, K. Estrada, and J. G. Tamez Pena, "Benchmarking machine learning models for late-onset alzheimer's disease prediction from genomic data," *BMC bioinformatics,* vol. 20, 2019.

[17] G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim, "Predicting Alzheimer's disease progression using multi-modal deep learning approach," *Scientific reports,* vol. 9, no. 1, pp. 1-12, 2019.

[18] T. Jo, K. Nho, P. Bice, and A. J. Saykin, "Deep learning-based identification of genetic variants: application to Alzheimer's disease classification," (in eng), *Brief Bioinform,* vol. 23, no. 2, Mar 10 2022, doi: 10.1093/bib/bbac022.

[19] C. Maj *et al.*, "Integration of machine learning methods to dissect genetically imputed transcriptomic profiles in alzheimer's disease," *Frontiers in genetics,* vol. 10, p. 726, 2019.

[20] A. S. Alatrany, A. J. Hussain, J. Mustafina, and D. Al-Jumeily, "Machine Learning Approaches and Applications in Genome Wide Association Study for Alzheimer's Disease: A Systematic Review," *IEEE Access,* vol. 10, pp. 62831-62847, 2022, doi: 10.1109/ACCESS.2022.3182543.

[21] H. Ahmed, H. Soliman, and M. Elmogy, "Early Detection of Alzheimer's Disease Based on Single Nucleotide Polymorphisms (SNPs) Analysis and Machine Learning Techniques," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 2020: IEEE, pp. 1-6.

[22] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "An efficient classifier for Alzheimer's disease genes identification," *Molecules,* vol. 23, no. 12, p. 3140, 2018.

[23] O. Weissbrod, D. Geiger, and S. Rosset, "Multikernel linear mixed models for complex phenotype prediction," *Genome research,* vol. 26, no. 7, pp. 969-979, 2016.

[24] Y. Wen and Q. Lu, "Multikernel linear mixed model with adaptive lasso for complex phenotype prediction," *Statistics in medicine,* vol. 39, no. 9, pp. 1311-1327, 2020.

[25] L. Liu, Q. Meng, C. Weng, Q. Lu, T. Wang, and Y. Wen, "Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data," *PLOS Computational Biology,* vol. 18, no. 7, p. e1010328, 2022.

[26] B. Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Frontiers in genetics,* vol. 10, p. 214, 2019.

[27] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, no. 10, pp. 1345-1359, 2010, doi: 10.1109/TKDE.2009.191.

[28] L. Koumakis, "Deep learning models in genomics; are we there yet?," *Computational and Structural Biotechnology Journal,* vol. 18, pp. 1466-1473, 2020.

[29] S. R. Dhruba, R. Rahman, K. Matlock, S. Ghosh, and R. Pal, "Application of transfer learning for cancer drug sensitivity prediction," *BMC bioinformatics,* vol. 19, no. 17, pp. 51-63, 2018.

[30] J. Singh, J. Hanson, K. Paliwal, and Y. Zhou, "RNA secondary

structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning," *Nature communications,* vol. 10, no. 1, pp. 1-13, 2019.

[31] Z. Zhao, L. G. Fritsche, J. A. Smith, B. Mukherjee, and S. Lee, "The construction of cross-population polygenic risk scores using transfer learning," *The American Journal of Human Genetics,* vol. 109, no. 11, pp. 1998-2008, 2022/11/03/ 2022, doi: https://doi.org/10.1016/j.ajhg.2022.09.010.

[32] B. Azarkhalili, A. Saberi, H. Chitsaz, and A. Sharifi-Zarchi, "DeePathology: Deep Multi-Task Learning for Inferring Molecular Pathology from Cancer Transcriptome," *Scientific Reports,* vol. 9, no. 1, p. 16526, 2019/11/11 2019, doi: 10.1038/s41598-019-52937-5.

[33] D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Research,* Article vol. 26, no. 7, pp. 990-999, 2016, doi: 10.1101/gr.200535.115.

[34] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics,* vol. 20, no. 7, pp. 389-403, 2019/07/01 2019, doi: 10.1038/s41576-019-0122-6.

[35] P. Tian, T. H. Chan, Y. F. Wang, W. Yang, G. Yin, and Y. D. Zhang, "Multiethnic polygenic risk prediction in diverse populations through transfer learning," (in eng), *Front Genet,* vol. 13, p. 906965, 2022, doi: 10.3389/fgene.2022.906965.

[36] M. Muneeb, S. Feng, and A. Henschel, "Transfer learning for genotype–phenotype prediction using deep learning models," *BMC Bioinformatics,* vol. 23, no. 1, p. 511, 2022/11/29 2022, doi: 10.1186/s12859-022-05036-8.

[37] S. Purcell *et al.*, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics,* vol. 81, no. 3, pp. 559-575, 2007/09/01/ 2007, doi: https://doi.org/10.1086/519795.

[38] J. A. Webster *et al.*, "Genetic control of human brain transcript expression in Alzheimer disease," *The American Journal of Human Genetics,* vol. 84, no. 4, pp. 445-458, 2009.

[39] A. Stella *et al.*, "AdaptMap: exploring goat diversity and adaptation," *Genetics Selection Evolution,* vol. 50, no. 1, p. 61, 2018/11/19 2018, doi: 10.1186/s12711-018-0427-5.

[40] L. Colli *et al.*, "Genome-wide SNP profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes," *Genetics Selection Evolution,* vol. 50, no. 1, p. 58, 2018/11/19 2018, doi: 10.1186/s12711-018-0422-x.

[41] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," (in eng), *Nat Protoc,* vol. 5, no. 9, pp. 1564-73, Sep 2010, doi: 10.1038/nprot.2010.116.

[42] N. Kleanthous, A. Hussain, W. Khan, J. Sneddon, and A. Mason, "Feature Extraction and Random Forest to Identify Sheep Behavior from Accelerometer Data," in *International Conference on Intelligent Computing*, 2020: Springer, pp. 408-419.

[43] F. Bertolini *et al.*, "Signatures of selection and environmental adaptation across the goat genome post-domestication," *Genetics Selection Evolution,* vol. 50, no. 1, p. 57, 2018/11/19 2018, doi: 10.1186/s12711-018-0421-y.

[44] L. Wang, Z.-H. You, Y.-A. Huang, D.-S. Huang, and K. C. Chan, "An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network," *Bioinformatics,* vol. 36, no. 13, pp. 4038-4046, 2020.

[45] L. Wang, Z. H. You, D. S. Huang, and F. Zhou, "Combining High Speed ELM Learning with a Deep Convolutional Neural Network Feature Encoding for Predicting Protein-RNA Interactions," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 17, no. 3, pp. 972-980, 2020, doi: 10.1109/TCBB.2018.2874267.

[46] Z. Shen, S. P. Deng, and D. S. Huang, "RNA-Protein Binding Sites Prediction via Multi Scale Convolutional Gated Recurrent Unit Networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 17, no. 5, pp. 1741-1750, 2020, doi: 10.1109/TCBB.2019.2910513.

[47] D. Delano, M. Eberle, L. Galver, and C. Rosenow, "Array differences in genomic coverage and data quality impact GWAS success," *Illumina,* 2010.

[48] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer genomics & proteomics,* vol. 15, no. 1, pp. 41-51, 2018.

[49] N. Kleanthous, A. Hussain, W. Khan, J. Sneddon, and P. Liatsis, "Deep transfer learning in sheep activity recognition using accelerometer data," *Expert Systems with Applications,* vol. 207, p. 117925, 2022/11/30/ 2022, doi: https://doi.org/10.1016/j.eswa.2022.117925.

[50] B. Yang *et al.*, "BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone," *Bioinformatics,* vol. 33, no. 13, pp. 1930-1936, 2017.

[51] Q. Liao, Y. Ding, Z. L. Jiang, X. Wang, C. Zhang, and Q. Zhang, "Multi-task deep convolutional neural network for cancer diagnosis," *Neurocomputing,* vol. 348, pp. 66-73, 2019.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[54] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods,* vol. 12, no. 10, pp. 931-934, 2015.

[55] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[56] L. Perreauls. "PyPlink." https://lemieuxl.github.io/pyplink/pyplink.html. (accessed 5/7/2022.

[57] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[58] Q. Zhang, L. Zhu, and D.-S. Huang, "High-order convolutional neural network architecture for predicting DNA-protein binding sites," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 16, no. 4, pp. 1184-1192, 2018.

[59] Z. Shen, S.-P. Deng, and D.-S. Huang, "Capsule network for predicting RNA-protein binding preferences using hybrid feature," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 17, no. 5, pp. 1488-1492, 2019.

[60] Q. Zhang, Z. Shen, and D.-S. Huang, "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network," *Scientific reports,* vol. 9, no. 1, pp. 1-12, 2019.

[61] H. Zhang, L. Zhu, and D. S. Huang, "DiscMLA: An Efficient Discriminative Motif Learning Algorithm over High-Throughput Datasets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 15, no. 6, pp. 1810-1820, 2018, doi: 10.1109/TCBB.2016.2561930.

[62] C. Peng, Y. Zheng, and D. S. Huang, "Capsule Network Based Modeling of Multi-omics Data for Discovery of Breast Cancer-Related Genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 17, no. 5, pp. 1605-1612, 2020, doi: 10.1109/TCBB.2019.2909905.

[63] L. Yuan *et al.*, "Integration of Multi-Omics Data for Gene Regulatory Network Inference and Application to Breast Cancer," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 16, no. 3, pp. 782-791, 2019, doi: 10.1109/TCBB.2018.2866836.

[64] M. Arnal Segura *et al.*, "Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring,* vol. 14, no. 1, p. e12300, 2022.

[65] M. R. Robinson, N. R. Wray, and P. M. Visscher, "Explaining additional genetic variation in complex traits," *Trends in Genetics,* vol. 30, no. 4, pp. 124-132, 2014.

[66] N. Raghavan and G. Tosto, "Genetics of Alzheimer’s disease: the importance of polygenic and epistatic components," *Current neurology and neuroscience reports,* vol. 17, no. 10, pp. 1-10, 2017.

[67] L. Zhu, S. Deng, Z. You, and D. Huang, "Identifying Spurious Interactions in the Protein-Protein Interaction Networks Using Local Similarity Preserving Embedding," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 14, no. 2, pp. 345-352, 2017, doi: 10.1109/TCBB.2015.2407393.

[68] W. Lee, D.-S. Huang, and K. Han, "Constructing cancer patient-specific and group-specific gene networks with multi-omics data," *BMC Medical Genomics,* vol. 13, no. 6, p. 81, 2020/08/27 2020, doi: 10.1186/s12920-020-00736-7.

[69] X. Liang, L. Zhu, and D. Huang, "Optimization of Gene Set Annotations Using Robust Trace-Norm Multitask Learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 15, no. 3, pp. 1016-1021, 2018, doi: 10.1109/TCBB.2017.2690427.

[70] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O’Sullivan, "Machine learning SNP based prediction for precision medicine," *Frontiers in genetics,* vol. 10, p. 267, 2019.

[71] A. B. Popejoy and S. M. Fullerton, "Genomics is failing on diversity," *Nature,* vol. 538, no. 7624, pp. 161-164, 2016.

[72] R. D. Dowell, "The similarity of gene expression between human and mouse tissues," *Genome Biology,* vol. 12, no. 1, p. 101, 2011/01/17 2011, doi: 10.1186/gb-2011-12-1-101.

[73] Y. Shi and D. M. Holtzman, "Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight," *Nature Reviews Immunology,* vol. 18, no. 12, pp. 759-772, 2018.

ABBAS S. ALATRANY received the BSc degree from Al-Mustansiriya University, Iraq, in 2017, the PG Diploma from University of Leicester, UK in 2020. Currently working towards the PhD degree in computer science from Liverpool John Moores University, U.K. His research interest includes computational biology and bioinformatics.

WASIQ KHAN (Member, IEEE) received his Ph.D. in AI and speech processing from Bradford University, U.K., in 2015. He is currently a Senior academic in Artificial Intelligence and data sciences with the Department of Computer Science, Liverpool John Moores University, U.K. He is research active within the domain of applied AI/machine learning, video/speech data processing, and data analytics. He has been working as a Lead on various large-scale research projects in collaborations with academia and industry mainly related to applied AI and pattern matching. Along with his Ph.D. supervisions and academic roles, he has established academic citizenship within the domain of AI and data science. He is also a fellow of HEA and a member of the Computational Intelligence Society

ABIR J. HUSSAIN (Member, IEEE) received the Ph.D. degree from The University of Manchester (UMIST), U.K., in 2000. Her Ph.D. thesis title Polynomial Neural Networks for Image and Signal Processing. She is currently a Professor in machine learning with the University of Sharjah, United Arab Emirates. She has published numerous refereed research papers in conferences and journals in the research areas of neural networks, signal prediction, telecommunication fraud detection, and image compression. She is one of the initiators and chairs of the Development in e-Systems Engineering (DeSE) series, most notably illustrated by the IEEE technically sponsored DeSE International Conference Series. Her research interests include machine learning algorithms and their applications to medical, image and signal processing, and data analysis.

JAMILA MUSTAFINA is a professor and a leader of iTech Research Lab at Kazan Federal University. She received her PhD in 2007 and full professorship in 2012. Her scientific interests cover the interdisciplinary approach towards the improvement of the social spheres of human life using technology and applied computing. She has published over 100 peer-reviewed scientific articles, 4 books, 8+ book chapters. Jamila has been awarded a number of research grants nationally and internationally. She is an International Committee Chair for IEEE International Conference Series on Developments in eSystems Engineering DeSE (www.dese.org.uk).

DHIYA AL-JUMEILY (Senior Member, IEEE) is currently a Professor in artificial intelligence with Liverpool John Moores University and the President of the e-Systems Engineering Society. He has extensive research interests include the wide variety of interdisciplinary perspectives concerning the theory and practice of applied artificial intelligence in medicine, human biology, intelligent community, and health care. He has published over 300 peer-reviewed scientific international publications. His current research passion is decision support systems for self-management of health and medicine. He is also a Chartered IT Professional. He is also a fellow of the U.K. Higher Education Academy. He is also a Senior Member of OBE.