2023

MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR GENE REGULATORY NETWORK INFERENCE IN PLANT SPECIES

Sai Teja Mummadi
Michigan Technological University, mummadi@mtu.edu

MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR GENE REGULATORY NETWORK INFERENCE IN PLANT SPECIES

By

Sai Teja Mummadi

A REPORT

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Computer Science

MICHIGAN TECHNOLOGICAL UNIVERSITY

2023

This report has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Computer Science.

Department of Computer Science

Report Advisor:	*Dr. Hairong Wei*
Committee Member:	*Dr. Dukka KC*
Committee Member:	*Dr. Xiaoyong (Brian) Yuan*
Department Chair:	*Dr. Andy Duan*

Table of Contents

List of Figures

List of Tables

Abstract

The construction of gene regulatory networks (GRNs) is vital for understanding the regulation of metabolic pathways, biological processes, and complex traits during plant growth and responses to environmental cues and stresses. The increasing availability of public databases has facilitated the development of numerous methods for inferring gene regulatory relationships between transcription factors and their targets. However, there is limited research on supervised learning techniques that utilize available regulatory relationships of plant species in public databases.

This study investigates the potential of machine learning (ML), deep learning (DL), and hybrid approaches for constructing GRNs in plant species, specifically *Arabidopsis thaliana*, poplar, and maize. Challenges arise due to limited training data for gene regulatory pairs, especially in less-studied species such as poplar and maize. Nonetheless, our results demonstrate that hybrid models integrating ML and artificial neural network (ANN) techniques significantly outperformed traditional methods in predicting gene regulatory relationships. The best-performing hybrid models achieved over 95% accuracy on holdout test datasets, surpassing traditional ML and ANN models and also showed good accuracy on lignin biosynthesis pathway analysis.

Employing transfer learning techniques, this study has also successfully transferred the known knowledge of gene regulation from one species to another, substantially improving performance and manifesting the viability of cross-species learning using deep learning-based approaches. This study contributes to the methodology for growing body of knowledge in GRN prediction and construction for plant species, highlighting the value of

adopting hybrid models and transfer learning techniques. This study and the results will help to pave a way for future research on how to learn from known to unknown and will be conductive to the advance of modern genomics and bioinformatics.

1 Introduction

A gene regulatory network (GRN) is a bipartite directional graph which depicts connections between regulators and their target genes. “Regulators” generally refers to components such as transcription factors (TFs) that regulate gene expression. “Target genes” generally refers to genes with non-regulatory functions, like structural genes, pathway genes, and signaling genes, as well as genes with regulatory functions such as TFs. A connection in a GRN, also known as an edge, represents the regulatory relationship between a regulator and a target gene. Regulators play important roles because their target genes may be involved in many metabolic pathways and biological processes that are the basis for the growth and development of organisms. A GRN is directional since regulators control target genes, but not the other way around.

In plants, GRNs are built to understand how different metabolic pathways, biological processes, and complex traits are regulated during growth and their role in responses to various environmental cues. Though GRNs can be constructed through conventional experimental approaches, including DNA electrophoretic mobility shift assay (EMSA) (Hellman and Fried 2007), a yeast one hybrid assay(Wilson et al. 1991), chromatin immunoprecipitation (ChIP) and DNA-sequencing (ChIP-seq) (Robertson et al. 2007), and DNA affinity purification and sequencing (DAP-seq) (Bartlett et al. 2017), these approaches are labor intensive and time consuming. As a result, the regulatory relationships inferred by such approaches are limited to a small number of genes. In contrast, techniques such as microarray and RNA-seq can generate terabytes of transcriptomic data for inferring gene networks. To leverage such readily available open-source gene transcriptome data, various algorithms and data analysis tools have been developed to construct GRNs.

2 Related Work

The approach to analyzing transcriptomic data depends on the nature of the data and the method used to generate it. Initially, transcriptomic data was generated mainly from single-celled organisms such as yeast or bacteria. These small organisms allowed researchers to produce time-course data in closely spaced intervals. This temporal data allowed for the effective use of dynamic algorithms such as finite state(Ruklisa, Brazma, and Viksna 2005) and dynamic Bayesian networks (Dojer et al. 2006), which require time-series data with small intervals to be able to successfully predict gene regulatory relationships. In contrast, it is complicated and difficult to generate time-course transcriptomic data (especially with small time intervals) from multicellular organisms such as plants and animals due to the time-consuming harvesting processes involved. As a result, most transcriptomic data from plants and animals is static data, produced from either a small-scale treatment versus control design or a time-series design with large time intervals (on the scale of days). A number of algorithms have been developed for the analysis of static data that do not rely on any temporal variables to simulate gene regulatory relationships. These static algorithms include Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE)(Margolin et al. 2006) and mutual information-based algorithms (Butte and Kohane 2000), (Luo, Hankenson, and Woolf 2008).

In the recent years, many algorithms have been developed and used for the construction of the hierarchical gene regulatory networks, such as backward elimination and Random Forest (BWERF) algorithm (Deng et al. 2017), Bottom-up Gaussian graphic Model (GGM) algorithm (Deng et al. 2017). Moreover, the algorithms that can be used to identify biological pathway regulators have been developed, such as TGMI (Gunasekara

et al. 2018) and Huber-Berhu-partial least squares (HB-PLS). Furthermore, there are several methods which is capable of constructing multiple GRNs using the data from multiple tissues or conditions, such as JRmGRN (Deng et al. 2018) and joint graphical lasso using ADMM (Danaher, Wang, and Witten 2014).

Machine Learning (ML) algorithms which learn from the available data are also used in modern biology and have advanced GRN predictions. A tree-based ML algorithm, GENIE3 (Huynh-Thu et al. 2010), has been widely used in recent GRN inference research and is applicable to both static and dynamic transcriptomic data. GENIE3 (Huynh-Thu et al. 2010) has been utilized with a number of publicly available datasets consisting of RNA-Seq data from various parts of the plant, including leaf, shoot, seed, etc. Supervised machine learning methods involve learning from labeled data to make predictions or classifications. These methods have not been extensively explored for gene expression data in plant species, primarily focusing on single-cell organisms and animal species. For instance, supervised learning techniques, such as multiple linear regression, Support Vector Machine (SVM), and Decision Trees, have been employed to reconstruct the budding yeast cell cycle network (Jochen Supper 2007).

Ensemble techniques, which combine multiple ML algorithms to enhance prediction performance, have also been used in related areas. For example, Ensemble methods for Gene Regulatory Networks using Topological features (EnGRNT) is an ensemble learning technique applied to drug design (B. Khojasteh 2021). However, research on supervised learning approaches for predicting GRNs using gene expression data specifically in plant species remains limited.

3 Multiple OMICS Data Collection

Transcriptomic data from three plant species, *Arabidopsis thaliana*, poplar, and maize, were collected from the Sequence Read Archive (SRA) database, National Center for Biotechnology Information (NCBI). The database includes numerous species and gene information, including expression data, sequencing data of all the genes in the genome of the particular species.

The data for *Arabidopsis* (Compendium Data Set 1) includes 22,093 genes and 1,253 biological samples. The samples were collected in different RNA-seq experiments conducted by different data submitters. The samples were collected from multiple tissues such as stems, shoots, roots, leaves, etc. The transcriptome data sets of maize and poplar were collected in a similar way. The poplar transcriptome data (Compendium Data Set 2) consists of 34,699 genes and 743 biological samples. The maize transcriptome data (Compendium Data Set 3) consists of 39,756 genes and 1,626 biological samples.

Table 3.1: The RNA-Seq data sets of Arabidopsis (Compendium Data Set 1), poplar (Compendium Data Set 2), and maize (Compendium Data Set 3) which were downloaded from Sequence Read Archive(SRA), NCBI (https://www.ncbi.nlm.nih.gov/sra).

	Arabidopsis **(Compendium Data Set 1)**	**Poplar (Compendium Data Set 2)**	**Maize (Compendium Data Set 3)**
Number of Genes	22,093	34,699	39,756
Expression Samples	1,253	743	1,626

3.1 *Arabidopsis* training data and testing data

The following section explains the process of extracting training data for *Arabidopsis* and the collection of the Transcriptomic Test Data Sets. The first part discusses the training data, and the second part discusses the testing data.

3.1.1 *Arabidopsis* training data

The 1,253 *Arabidopsis* RNA-seq transcriptomic data sets were downloaded from the Sequence Read Archive (SRA), NCBI database (Compendium Data Set 1) and processed as follows: Adaptor sequences and low-quality bases of raw reads were trimmed by Trimmomatic (version 0.38) (Bolger, Lohse, and Usadel 2014), a software developed for removing adapter sequences and also for filtering low quality reads. After trimming the sequence based on the sliding window, the cleaned reads are more effective for further analysis. Trimmed reads were aligned to the *Arabidopsis* TAIR10 reference genome using STAR (2.7.3a) (Dobin et al. 2013). Uniquely mapped reads were used for counting reads per gene.

Raw counts were normalized with the weighted trimmed mean of M-values (TMM) algorithm contained in edgeR (Robinson, McCarthy, and Smyth 2010). After the preprocessing described above, Compendium Data Set 1 consisted of 1,253 samples, as shown in Table 3.1. To obtain the training data set from Compendium Data Set 1, the 1,231 pairs of positive regulatory relationships proven by experimental validation were extracted from the regulatory relationships curated in the *Arabidopsis* Gene Regulatory Information Server (AGRIS) database (Yilmaz et al. 2011). In addition, 1,231 pairs of putative negative gene pairs were obtained by randomly combining the TFs from the positive regulatory

relationships with other genes in the genome. If a combination was coincidently a positive pair, it was discarded until a total of 1,231 pairs of negative regulatory relationships of genes were obtained. Combinations of positive and negative pairs and their expression values across 1,253 samples in Compendium Data Set 1 resulted in a data matrix with 2,462 rows × 2,511 columns. In this data matrix, five columns are related to the TF gene id and target gene id along with the true label for the data, and the other columns denote the RNA-seq data for a sample; each row denotes the transcriptomic data of 1,253 samples of TF and target genes. The TFs and targets are juxtaposed in the data set; therefore, the first half of the columns denotes the TF and the other half denotes the target genes.

Table 3.2: *Arabidopsis* training data set. The positive regulatory relationships of the 1,231 pairs shown in this table were downloaded from the Arabidopsis Gene Regulatory Information Server (AGRIS) database (Yilmaz et al. 2011). The negative regulatory relationships of the 1,231 pairs shown in this table were a random combination of genes that are not shown as positive regulatory relationships in AGRIS. The gene expression data (RNA-seq) used was from 1,253 multi-tissue samples of Arabidopsis downloaded from the Sequence Read Archive (SRA), NCBI database (referred to as Compendium Data Set 1 in this Report).

	TF	Gene	Regulation	TF.1	ERR2352507	ERR2352508	ERR2352509	ERR2352510	SRR10060476	SRR10060477
0	AT5G16560	AT1G01140	True	AT5G16560	52.311045	67.884831	64.499893	62.119688	187.071209	181.071445
1	AT2G20180	AT1G01260	True	AT2G20180	448.606837	452.112976	490.499184	492.356042	333.123481	329.740631
2	AT5G16560	AT1G01490	True	AT5G16560	52.311045	67.884831	64.499893	62.119688	187.071209	181.071445
3	AT2G20180	AT1G01650	True	AT2G20180	448.606837	452.112976	490.499184	492.356042	333.123481	329.740631
4	AT2G20180	AT1G01720	True	AT2G20180	448.606837	452.112976	490.499184	492.356042	333.123481	329.740631
...	...	...	...	...	...	...	...	...	...	...
1226	AT1G32770	AT5G67230	True	AT1G32770	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1227	AT1G71930	AT5G67230	True	AT1G71930	12.681465	12.219270	8.999985	1.150365	14.915977	13.977445
1228	AT3G08500	AT5G67230	True	AT3G08500	17.437015	12.219270	11.999980	11.503646	12.429981	10.165414
1229	AT5G12870	AT5G67230	True	AT5G12870	30.118480	17.650056	17.999970	16.105104	29.831954	28.590228
1230	AT5G16560	AT5G67440	True	AT5G16560	52.311045	67.884831	64.499893	62.119688	187.071209	181.071445
0	AT4G17570	AT1G01130	False	AT4G17570	217.170094	175.142865	196.499673	207.065625	270.352079	277.007544
1	AT3G14180	AT1G01200	False	AT3G14180	432.755005	454.828370	440.999266	475.100573	467.367272	432.665452
2	AT3G19290	AT1G01370	False	AT3G19290	578.591857	492.843875	493.499179	607.392500	352.389951	358.966198
3	AT2G20400	AT1G01420	False	AT2G20400	103.036906	95.038764	95.999840	98.931354	86.388365	72.428578
4	AT5G16570	AT1G01440	False	AT5G16570	19.022198	10.861573	5.999990	12.654010	54.070416	62.898502
...	...	...	...	...	...	...	...	...	...	...
1226	AT3G11280	ATCG01060	False	AT3G11280	421.658723	442.609100	443.999261	448.642188	625.228026	590.864715
1227	AT1G75550	ATMG00070	False	AT1G75550	44.385129	42.088595	47.999920	40.262760	37.911441	41.296996
1228	AT1G51950	ATMG00660	False	AT1G51950	507.258614	487.413089	452.999246	599.339948	290.240048	273.830852
1229	AT3G46090	ATMG00665	False	AT3G46090	25.362931	14.934663	35.999940	35.661302	39.775938	15.883460
1230	AT2G36400	AT1G01100	False	AT2G36400	293.258886	304.124044	241.499598	364.665573	435.049322	426.947407

In Table 3.2, the overall layout of the training data with positive and negative relationships is evident. When the genes in this training data set are examined, there are 432 unique TF genes in the data, 1,833 unique target genes, and 79 of the genes are both TFs as well as target genes. These complex relationships can be visualized as shown in Figure 3.1, where green nodes denote target genes, red nodes denote TF genes, and blue nodes are TFs that are also regulated by other TFs. The network is largely a scale-free gene regulatory network (GRN) that has been observed in many species (Dewey GT 2000-2013).

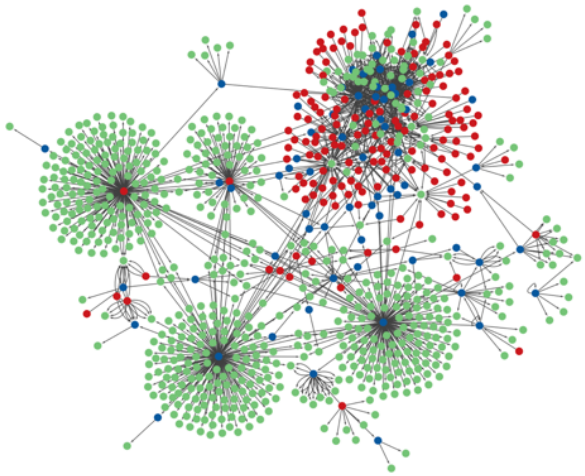

Figure 3.1: Gene regulatory network (GRN) that represents the *Arabidopsis* training data, also referred to as Transcriptomic Training Data Set 1. Green nodes denote the target genes, red nodes denote the regulators, and blue nodes are transcription factors (TFs) that are regulated by other TFs. Note that only positive regulatory relationships are shown.

3.1.2 *Arabidopsis* testing data

In this report, two testing data for *Arabidopsis* species were used. *Arabidopsis* Transcriptomic Test Data Set 1, *Arabidopsis* Transcriptomic Test Data Set 2.

*Arabidopsi*s Transcriptomic Test Data Set 1 consists of genes associated with the lignin biosynthesis pathway (LBP) (Boerjan, Ralph, and Baucher 2003), a metabolic pathway involved in plant growth and development through the production of lignin, a major component of the cell wall. LBP genes include a few dozen genes whose proteins function as enzymes to catalyze a series of biochemical reactions that lead to the synthesis of three monolignols, p-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol, which are then polymerized by laccases. LBP genes include phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), coumaroyl-CoA ligase (4CL), hydroxycinnamoyltransferase (HCT), caffeoyl-CoA O-methyltransferase (CCoAOMT), p-coumarate 3-hydroxylase (C3H), ferulate 5-hydroxylase (F5H), caffeic acid O-

methyltransferase (COMT), cinnamyl alcohol dehydrogenase(CAD), laccase (LAC), cinnamoyl-CoA reductase (CCR), and peroxidase (POD). To identify potential regulatory genes of LBP, 1,415 unique TFs were collected from the AGRIS (Yilmaz et al. 2011). The TFs are paired with 20 target genes from the lignin pathway which resulted in 28,300 pairs for Arabidopsis Transcriptomic Test Data Set 1. The expression data of 28,300 pairs were extracted from Compendium Data Set 1 with 1,253 samples.

Arabidopsis Transcriptomic Test Data Set 2 incorporates data from Taylor-Teeples's Supplementary Table 2 (Taylor-Teeples et al. 2015). This data set comprises 199 transcription factors (TFs) and 35 target genes, with 582 regulatory pairs which are considered as true regulatory relationships because they were validated by using Yeast One Hybrid System (Bulyk et al. 1999). These relationships were confirmed through functional enrichment analysis and comparison with existing literature. Furthermore, a negative testing data set was generated by randomly pairing genes that were not identified as positive regulatory pairs in the AGRIS (Yilmaz et al. 2011). The 582 positive regulatory pairs were combined with 582 randomly generated negative pairs, resulting in a total of 1,164 regulatory pairs in *Arabidopsis* Transcriptomic Test Data Set 2. Expression data for these 1,164 pairs were extracted from Compendium Data Set 1, which contains 1,253 samples.

3.2 Poplar training data and testing data

The following section explains the process of extracting training data for the poplar species and the collection of the Transcriptomic Test Data Sets. The first part discusses the training data, and the second part discusses the testing data.

3.2.1 Poplar training data

The RNA-seq transcriptomic data sets from 743 samples were downloaded from the Sequence Read Archive (SRA), NCBI database and processed as described for *Arabidopsis*. Trimmed reads were aligned to the poplar reference genome using STAR (2.7.3a). Uniquely mapped reads were used for counting reads per gene.

The raw counts were normalized with weighted trimmed mean of M-values (TMM) algorithm contained in edgeR (Robinson, McCarthy, and Smyth 2010). After above preprocessing, I eventually obtained a compendium data set, referred to as Compendium Data Set 2, which contains 743 samples, as shown in Table 3.1. The 2,107 pairs of TFs and their putative targeted gene in poplar obtained by homologous mapping of *Arabidopsis* experimented verified regulatory gene pairs from the AGRIS (Yilmaz et al. 2011) 2011) were used as the positive training data set. Meanwhile, the 2,107 pairs of putative negative gene pairs were obtained by a random combination of the TFs from the positive regulatory to the other genes in the genome. If a combination is coincidently a positive pair, discard the pair until a total of 2,107 pairs of negative regulatory relationships of genes are obtained. Combination of positive and negative pairs and their expression values across 743 samples in Compendium Data Set 2 result in a data matrix with 4,214 rows × 1,491 columns. In this data matrix, 5 columns are related to the TF gene id and target gene id along with the true label for the data and the other columns denotes an RNA-seq data of a sample, and each row denotes the transcriptomic data of 743 samples of TF and Target genes. The TFs and targets are juxtaposed in the data set; therefore, the first half columns denote the TF and next half denotes the target genes.

Table 3.3: Poplar training data set. The positive regulatory relationships of 2,107 pairs shown in this table were obtained by homologous mapping of *Arabidopsis* regulatory pairs shown in *Arabidopsis* training data. The negative regulatory relationships of the 2,107 pairs shown in this table were a random combination of genes not shown in the positive relationships. The gene expression data used was RNA-Seq data from 743 samples downloaded from the SRA, NCBI database (Compendium Data Set 2).

	TF	Target	Regulation	TF.1	ERR1864428	ERR1864429	ERR1864430	ERR1864431	ERR1864432	ERR1864433
0	Potri.001G026000	Potri.013G113100	True	Potri.001G026000	463.606940	516.402571	310.957126	468.802316	360.542991	323.235742
1	Potri.001G026000	Potri.019G083600	True	Potri.001G026000	463.606940	516.402571	310.957126	468.802316	360.542991	323.235742
2	Potri.003G200000	Potri.013G113100	True	Potri.003G200000	130.294761	132.433706	97.957190	140.414220	135.976214	134.959253
3	Potri.003G200000	Potri.019G083600	True	Potri.003G200000	130.294761	132.433706	97.957190	140.414220	135.976214	134.959253
4	Potri.T126206	Potri.013G113100	True	Potri.T126206	118.174318	95.103399	62.647040	106.443038	76.229090	89.972835
...	...	...	...	...	...	...	...	...	...	...
2102	Potri.013G157900	Potri.004G174400	True	Potri.013G157900	2451.359571	565.287496	433.973133	1073.489361	358.482745	238.261397
2103	Potri.013G157900	Potri.009G134000	True	Potri.013G157900	2451.359571	565.287496	433.973133	1073.489361	358.482745	238.261397
2104	Potri.019G130700	Potri.004G174400	True	Potri.019G130700	2996.779500	1511.877406	1312.170731	4566.859276	828.218757	758.104446
2105	Potri.019G130700	Potri.009G134000	True	Potri.019G130700	2996.779500	1511.877406	1312.170731	4566.859276	828.218757	758.104446
2106	Potri.015G061900	Potri.006G213500	True	Potri.015G061900	1596.868348	3564.155436	3273.592606	3401.647723	1716.184637	1712.816200
0	Potri.002G019900	Potri.017G053400	False	Potri.002G019900	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Potri.003G103300	Potri.006G173700	False	Potri.003G103300	10472.062640	7226.080722	6508.457950	7259.641661	10548.457790	8629.061380
2	Potri.008G110800	Potri.002G149000	False	Potri.008G110800	490.877936	241.758174	132.128303	270.637086	55.626633	21.660127
3	Potri.001G208200	Potri.016G113700	False	Potri.001G208200	1133.261408	1616.757790	2093.550177	1415.465930	1318.557224	1516.208893
4	Potri.002G150000	Potri.003G181200	False	Potri.002G150000	251.499190	712.831087	717.593369	733.777538	197.783584	251.590706
...	...	...	...	...	...	...	...	...	...	...
2102	Potri.004G038200	Potri.001G049400	False	Potri.004G038200	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2103	Potri.012G141200	Potri.003G079400	False	Potri.012G141200	0.000000	3.555267	2.278074	0.000000	0.000000	0.000000
2104	Potri.004G073000	Potri.009G147900	False	Potri.004G073000	518.148933	3107.303592	2721.159616	2510.470373	3143.934882	3712.212542
2105	Potri.002G072800	Potri.010G216900	False	Potri.002G072800	0.000000	0.000000	0.000000	0.000000	0.000000	1.666164
2106	Potri.009G065300	Potri.019G012200	False	Potri.009G065300	1372.640155	1676.308517	1710.833714	1681.573525	1650.256776	1682.825255

In this poplar training data set, there are 962 unique genes of which 204 genes are TFs and 47 genes are both TFs as well as target genes. These complex relations can be visualized with the help of the graph in Figure 3.2, in which green nodes denote target genes, red nodes denote TF genes. The blue are the common genes which are TFs, but they need to be regulated by other TFs. This will make them a target for the other TFs. Therefore, there are complex relations in the gene regulatory networks in which the chain of regulation is extended several levels.

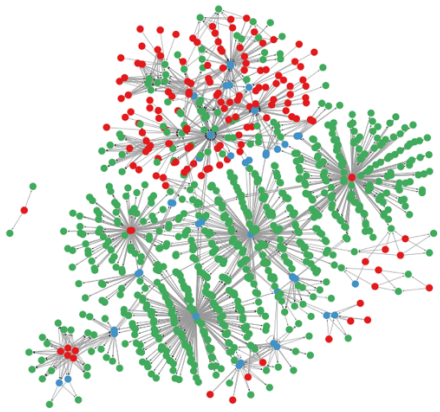

Figure 3.2: Gene regulatory network that represents the poplar training data. Green nodes denote the target genes, red nodes denote the regulators, and blue nodes are transcription factors (TFs) that are regulated by other TFs. Only positive regulations are shown.

3.2.2 Poplar testing data

The poplar Transcriptomic Test Data Set was prepared using genes related to the LBP. A total of 25 target genes for poplar species were identified through homologous mapping of *Arabidopsis* lignin pathway target genes, and 1,717 unique TFs were also found through homologous mapping of *Arabidopsis* TFs to poplar. Pairing 25 target genes with 1,717 unique TFs resulted in a total of 42,925 regulatory pairs in the poplar Transcriptomic Test Data Set. The gene expression data for 743 samples for the poplar Transcriptomic Test Data Set were extracted from the SRA database, NCBI (Compendium Data Set 2).

3.3 Maize training and testing data

The following section explains the process of extracting training data for the maize species and the collection of the Transcriptomic Test Data Sets. The first part discusses the training data, and the second part discusses the testing data.

3.3.1 Maize training data

The maize training data consists of 1,626 B73 maize paired-end RNA-Seq data from diverse tissues and treatments that were collected from the NCBI Sequence Read Archive

(SRA) database. The software Trimmomatic (version 0.38) (Bolger, Lohse, and Usadel 2014) was used to trim the adaptor sequences and low-quality bases of raw reads. Remaining paired-end reads were aligned to the B73 maize reference genome (B73Ref4) (Jiao et al. 2017) using STAR (version 2.6.0) (Dobin et al. 2013). Raw read counts per gene were calculated by STAR and then normalized by the library sizes of RNA-Seq samples to represent gene expression, resulting in a compendium data set, hereafter referred to as Compendium Data Set 3. 8,450 pairs of TFs and their putative target genes in maize were obtained by homologous mapping of experimentally verified *Arabidopsis* regulatory gene pairs from the AGRIS (Yilmaz et al. 2011) and were used as the positive training data set for deep learning.

Meanwhile, 8,450 pairs of putative negative gene pairs were obtained by randomly combining TFs from the positive regulatory relationships with other genes in the genome. If a combination was coincidently a positive pair, this pair was discarded until a total of 8,450 pairs of negative regulatory relationships of genes were obtained. The combination of positive and negative pairs and their expression values across 1,626 samples in Compendium Data Set 3 resulted in a data matrix with 16,900 rows × 3,257 columns. In this data matrix, five columns are related to the TF gene id and target gene id along with the true label for the data and the other columns denote the RNA-seq data for a sample; each row denotes the transcriptomic data of 1,626 samples of TFs and target genes. TFs and target genes are juxtaposed in the data set; therefore, the first half of the columns denotes the TFs and the other half denotes the target genes.

Table 3.4: Maize training data set. The positive regulatory relationships of 8,450 pairs shown in this table were obtained by homologous mapping of *Arabidopsis* regulatory

pairs shown in *Arabidopsis* training data. The negative regulatory relationships of the 8,450 pairs shown in this table were a random combination of genes not found in the positive relationships. The gene expression data was compendium data from 1,626 samples of B73 cultivar paired-end RNA-Seq data downloaded from the SRA, NCBI database (Compendium Data Set 3).

	TF	Target	Regulation	TF.1	ERR1314944	ERR1314945	ERR1314946	ERR1314947	ERR1314948	ERR1314949
0	Zm00001eb311960	Zm00001eb158010	True	Zm00001eb311960	24	24	30	98	114	98
1	Zm00001eb311960	Zm00001eb389330	True	Zm00001eb311960	24	24	30	98	114	98
2	Zm00001eb311960	Zm00001eb403620	True	Zm00001eb311960	24	24	30	98	114	98
3	Zm00001eb311960	Zm00001eb001940	True	Zm00001eb311960	24	24	30	98	114	98
4	Zm00001eb050790	Zm00001eb155400	True	Zm00001eb050790	39	25	30	910	1173	1063
...	...	...	...	...	...	...	...	...	...	...
8445	Zm00001eb410950	Zm00001eb273460	True	Zm00001eb410950	0	0	0	0	0	0
8446	Zm00001eb410950	Zm00001eb432200	True	Zm00001eb410950	0	0	0	0	0	0
8447	Zm00001eb093920	Zm00001eb068390	True	Zm00001eb093920	0	0	0	0	0	0
8448	Zm00001eb093920	Zm00001eb432200	True	Zm00001eb093920	0	0	0	0	0	0
8449	Zm00001eb093920	Zm00001eb225770	True	Zm00001eb093920	0	0	0	0	0	0
0	Zm00001eb216890	Zm00001eb421590	False	Zm00001eb216890	5	10	13	0	1	1
1	Zm00001eb133410	Zm00001eb046760	False	Zm00001eb133410	0	0	0	0	0	0
2	Zm00001eb038740	Zm00001eb345170	False	Zm00001eb038740	0	0	0	0	0	0
3	Zm00001eb389010	Zm00001eb169620	False	Zm00001eb389010	0	0	0	0	0	0
4	Zm00001eb406840	Zm00001eb169650	False	Zm00001eb406840	125	154	161	1710	1780	1516
...	...	...	...	...	...	...	...	...	...	...
8445	Zm00001eb410440	Zm00001eb184720	False	Zm00001eb410440	223	269	267	88	103	106
8446	Zm00001eb394380	Zm00001eb209640	False	Zm00001eb394380	31	28	28	22	31	39
8447	Zm00001eb024270	Zm00001eb290420	False	Zm00001eb024270	218	186	203	194	228	242
8448	Zm00001eb030870	Zm00001eb387410	False	Zm00001eb030870	1500	1500	1653	835	896	799
8449	Zm00001eb430750	Zm00001eb410980	False	Zm00001eb430750	497	515	375	2767	2626	2390

In this maize training data set, there are 1598 unique genes in which 501 genes are TFs and 115 genes are both TFs as well as target genes. These complex relations can be visualized as in the graph in Figure 3.3, in which green nodes denote target genes, red nodes denote TF genes. The blue are the common genes which are TFs, but they need to be regulated by other TF. As in poplar and *Arabidopsis*, there are complex relations in GRNs in which the chain of regulation is extended to several levels.

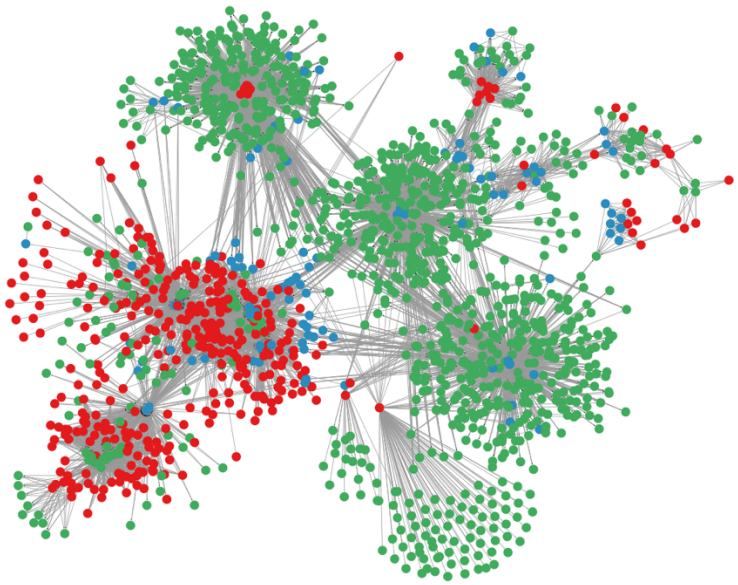

Figure 3.3: Gene regulatory network that represents the maize training data. Green nodes denote the target genes, red nodes denote the regulators, and blue nodes are transcription factors (TFs) that are regulated by other TFs. Only positive regulations are shown.

3.3.2 Maize testing data

The maize Transcriptomic Test Data Set was prepared with genes related to the LBP. Homologous mapping of the *Arabidopsis* lignin pathway target genes in maize resulted in a total of 38 target genes for the maize species. Pairing these 38 target genes with 2,555 unique TFs resulted in a total of 97,090 regulatory pairs in the maize Transcriptomic Test Data Set. The gene expression data of 1,626 samples for the maize Transcriptomic Test Data Set was extracted from Compendium Data Set 3.

3.4 Summary of all training and testing data

After above data collection and preprocessing, three distinct training datasets corresponding to the species *Arabidopsis*, poplar, and maize are presented in Table 3.5. The testing datasets for these species can be found in Table 3.6. Table 3.5 outlines total data points per species, the quantity of expression samples for each gene, and the counts of

both positive and negative pairs for every species. In Table 3.6, the testing data is displayed, along with the overall number of rows and the expression samples per gene.

Table 3.5: Distribution of *Arabidopsis*, poplar, and maize training data. *Arabidopsis*, poplar, and maize training data sets were extracted from Compendium Data Set 1 (*Arabidopsis*), Compendium Data Set 2 (poplar), and Compendium Data Set 3 (maize) as shown in Table 3.1.

Species	**Total Rows**	**Expression Samples**	**Positive Pairs**	**Negative Pairs**
Arabidopsis Training data	2,462	1,253	1,231	1,231
Poplar Training data	4,214	743	2,107	2,107
Maize Training data	16,900	1,626	8,450	8,450

Table 3.6: Distribution of *Arabidopsis*, poplar, and maize Transcriptomic Test Data Sets. The Transcriptomic Test Data Sets were extracted from Compendium Data Set 1 (*Arabidopsis*), Compendium Data Set 2 (poplar), and Compendium Data Set 3 (maize) as shown in Table 3.1.

Species	**Total Rows**	**Expression Samples**
Arabidopsis Transcriptomic Test Data Set 1	28,300	1,253
Arabidopsis Transcriptomic Test Data Set 2	1,164	1,253
Poplar Transcriptomic Test Data Set	42,925	743
Maize Transcriptomic Test Data Set	97,090	1,626

3.5 Data Cleaning and Transformation

After the data was collected, it was cleaned and transformed. Data cleaning is an important step in data preprocessing. The first step of data cleaning was removing additional columns such as gene IDs and true labels. Each data set downloaded from different sources had associated metadata. This metadata was removed, preserving only TF and target gene names and regulation information. Furthermore, the data had some missing values (e.g., NaN). NaN values were filled using various techniques such as Mean of the columns or Median of the columns. The gene data, however, was not related to the other rows and was truly independent of its neighbors; the NaN values in the gene expression

data were filled with 0 with the pandas fillna() function in Python. In addition, some of the data had redundant rows, which were removed from the data sets using the drop_duplicates() function from the pandas library. After the data was cleaned, it was normalized using the Z-score normalization technique with the help of standard scaler from the sklearn library. This technique transformed the data such that the mean value was 0 and the standard deviation of the data was 1. The same cleaning and transformation methods were applied to the test data sets for *Arabidopsis*, poplar, and maize.

Finally, the original training data was split into two different data sets: training data (80%) and validation data (20%). To ensure that both training, validation, and testing data had equal ratios of positive and negative relationships, stratified splitting was used along with data shuffling.

4 Methods

In this section, the methods of machine learning (ML) and deep learning that were employed to analyze the above gene expression data for predicting gene regulatory relationships and GRNs are recapitulated.

4.1 Machine Learning Models

ML models are algorithms that allow computer systems to learn and enhance their performance automatically by utilizing prior experience or existing data without explicit programming. These algorithms learn from the data on hand, recognize patterns, and make decisions without external input. Generally, ML models are trained on extensive data sets.

Supervised ML models, a category of ML techniques, require each data point to have a specific label. In binary classification problems, the label is either true or false. Supervised ML models, such as Support Vector Machines, Decision Trees, Logistic Regression, and K-Nearest Neighbors have demonstrated strong performance in inferring gene regulatory network (Jochen Supper 2007; Parry et al. 2010; Gillani et al. 2014; Choi et al. 2017). Support Vector Machines (Cortes 1995) are robust ML algorithms that identify hyperplanes to separate classes, maximizing the margin between points and the hyperplane. Decision Trees (Wu et al. 2008) employ a tree structure in which nodes represent attributes and leaves represent classes. The algorithm learns which attributes and values to use for splitting. Logistic Regression (McCullagh 1989) is a statistical model commonly utilized for classification problems. It produces an output ranging between 0 and 1, representing the probability of a particular class. The K-Nearest Neighbor (KNN) algorithm (Wu et al.

2008), one of the simplest ML methods, stores all data points, and when given a new point, assigns it to the class of its most similar neighbors.

Ensemble learning methods are supervised learning techniques which combine predictions from multiple base models, have also shown promising results in gene regulatory network prediction (Sergio Peignier 2021). Several ensemble techniques were implemented, including Random Forests, Extremely Randomized Trees, AdaBoost Models, Gradient Boosting, and Bagging. Random Forest classifier (Ho 1995) combines multiple Decision Trees, and its output is the class chosen by the majority of trees. It can also be used for regression problems, providing the mean of the Decision Trees. Extremely Randomized Trees (Geurts, Ernst, and Wehenkel 2006) which are also referred to as Extra Trees are a variation of Random Forest classifiers, where Decision Tree construction differs by randomly splitting features, increasing the model's variance and improving performance on unseen data. AdaBoost, short for Adaptive Booting (Wu et al. 2008) iteratively learns from misclassifications in previous iterations, using Decision Trees as its base model and assigning weights to samples based on classification accuracy. The Gradient Boosting algorithm (Friedman 2001) is akin to AdaBoost, employing iterative learning but training on mis-predicted samples. Lastly, the Bagging Classifier (Breiman 1996) is an ensemble technique in which individual models are trained on random subsets of the data, and their predictions are aggregated to produce the final prediction

4.2 Neural Networks

Neural networks, also known as artificial neural networks (ANNs), are a category of ML algorithms inspired by biological neurons. The fundamental element in an ANN is a

single neuron that receives input, performs computations, and passes the output to the next neuron. ANNs consist of several layers arranged sequentially, with each layer containing multiple neurons. Various types of ANNs exist, such as fully connected networks (FCNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). Neural networks have demonstrated promising results in classifying cancer using gene expression data (Rukhsar et al. 2022) and capturing gene expression relationships (Eetemadi and Tagkopoulos 2019). Additionally, CNN models have been employed to analyze microarray gene expression data (Tabares-Soto et al. 2020).

ANNs are trained using supervised learning with labeled inputs. Back propagation (Kelley 1960) optimizes weights and biases by minimizing prediction error using loss functions. Loss functions are crucial for building neural networks as they measure model performance on specific tasks by quantifying the error between predictions and actual outputs. Common loss functions include binary cross-entropy (BCE), hinge loss, mean squared error (MSE), mean squared logarithmic error (MSLE), mean absolute error (MAE), Poisson loss, Huber loss, and LogCosh loss depending on the type of problem. BCE is primarily used for classification tasks with binary target values. Hinge loss, an alternative to BCE, is employed in Support Vector Machine algorithms and works with output labels in the set of -1 and 1. MSE calculates the average squared difference between predictions and true labels, while MSLE first calculates the natural logarithm of the predicted values and then calculates the average of squared difference. MAE computes the average absolute difference between predictions and targets. Poisson loss, based on the Poisson distribution, calculates the negative log-likelihood, and can replace BCE loss in classification problems. Huber loss, which is dependent on MAE, provides a smooth

approximation and is more robust to outliers. LogCosh loss, though similar to MSE, is more robust to outliers. In this study, various loss functions are utilized in neural networks, and their predictions are analyzed.

4.2.1 Fully Connected Neural Networks

Fully connected networks are versatile and can be applied to a variety of problems. FCNs consist of input, hidden, and output layers. The hidden layers in the network consist of neurons which compute the weighted sum of previous inputs and apply an activation function, such as rectified linear unit (ReLU), hyperbolic tangent (tanh), or sigmoid activations. These functions introduce non-linearity, capturing complex relationships in data. Additional hidden layers, such as batch normalization and dropout, stabilize activations and prevent overfitting. Batch normalization layers normalize previous layer outputs in batches, stabilizing activations and reducing variance. Dropout layers prevent overfitting by setting a percentage of input neurons to zero during training, simulating various architectures. ReLU and sigmoid activations are used in dense and final layers, respectively, with sigmoid being ideal for binary classification problems. Figure 4.1 illustrates the general architecture of fully connected networks.

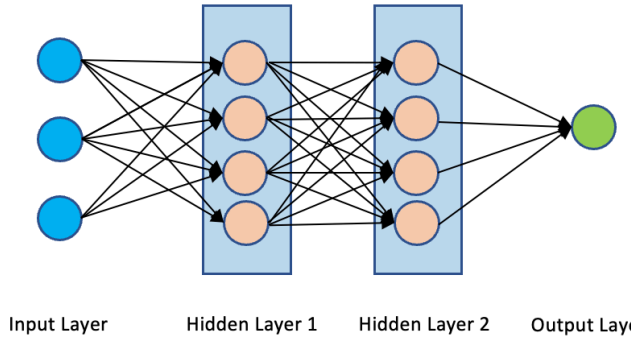

Figure 4.1: General architecture of fully connected neural networks. Blue nodes represent the input layer of fully connected neural networks, orange nodes denote the neurons in the hidden layers, and the green node is the output layer of the neural network.

4.2.2 Convolutional Neural Networks

CNNs are deep learning algorithms primarily used for image classification tasks. CNNs are adept at detecting spatial patterns in data, making them well-suited for gene expression analysis. They learn the significance of various features in an image using weights and biases. Convolution operations in CNNs reduce image size without losing essential characteristics, capturing temporal and spatial dependencies through relevant filters. Convolution mainly relies on kernel size, a fixed-size matrix that applies dot products sequentially over the image.

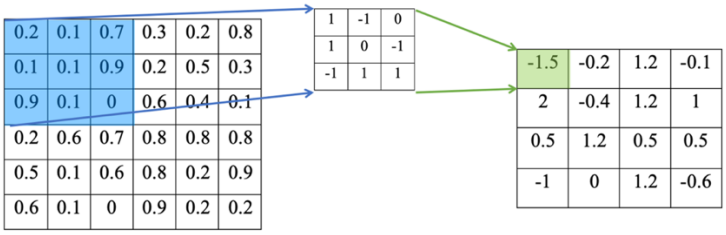

Figure 4.2: Example of convolution operation on a 6 x 6 matrix using a 3 x 3 kernel. This kernel is applied to the input matrix sequentially and the dot products are computed.

In Figure 4.2, the convolution operation is applied on a 6 x 6 square matrix using a 3 x 3 kernel. The output of a CNN layer, called a feature map, results from multiple convolutional layers applied sequentially. Another crucial operation in CNNs is pooling, which reduces feature map size without losing important data. A CNN comprises a set of convolutional layers followed by max pooling layers. After convolution, the data is flattened and fed into a regular neural network for classification. For the current study, the ReLU activation function and he-uniform kernel initialization were used (Sun 2015).

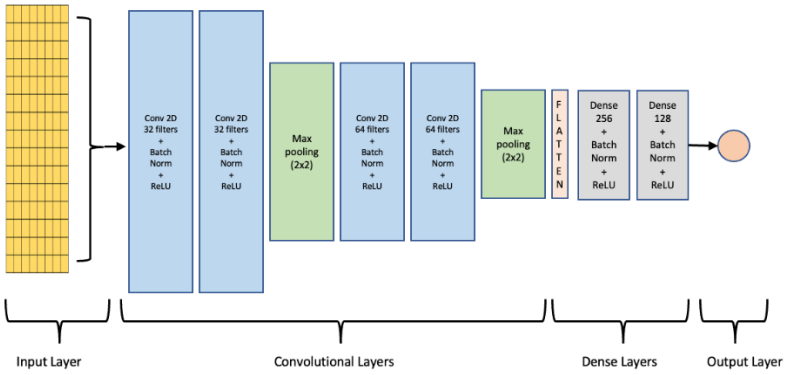

Figure 4.3: General architecture of Convolutional Neural Networks. The input layer is followed by a series of convolutional layers and max pooling layers. The output from the convolutional layers is flattened and fed as an input to the dense layers for classification. The output layer gives the probabilities of the classification.

In addition to the general CNNs I have also applied deep CNNs which involve skip connections such as ResNet Model (He et al. 2015), Mobile Net Model (Howard et al. 2017) and trained on the *Arabidopsis* data, poplar data and maize data.

4.3 Hybrid Architecture

Hybrid architecture combines neural networks and ML algorithms, leveraging the strengths of both approaches for effective gene expression data classification (Kong and Yu 2018). Our hybrid architecture consists of two steps: the feature extractor, also known as the convolutional encoder, and the classification model.

Hybrid architecture, Step 1: Convolutional encoder

The convolutional encoder plays a critical role in the hybrid architecture. As shown in Figure 4.4, input data is first passed through a series of convolutional and max-pooling layers, followed by dense layers for training and classification. Convolutional layers serve as feature extractors, while dense layers function as classification models. Outputs from

the convolutional layers are flattened to be fed into the dense layers. After successful training, the model is stored along with the learned weights.

Hybrid architecture, Step 2: Classification Model

In this step, outputs from the convolutional encoder/feature extractor are used to train ML models. Supervised ML Models such as Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors, as well as ensemble techniques like Random Forest, Extremely Randomized Trees, AdaBoost, Gradient Boosting, and Bagging are trained with the newly extracted features. CNN models excel in reducing data size while capturing essential information. Consequently, the CNN feature extractor is combined with traditional ML models to enhance gene regulatory network prediction accuracy.

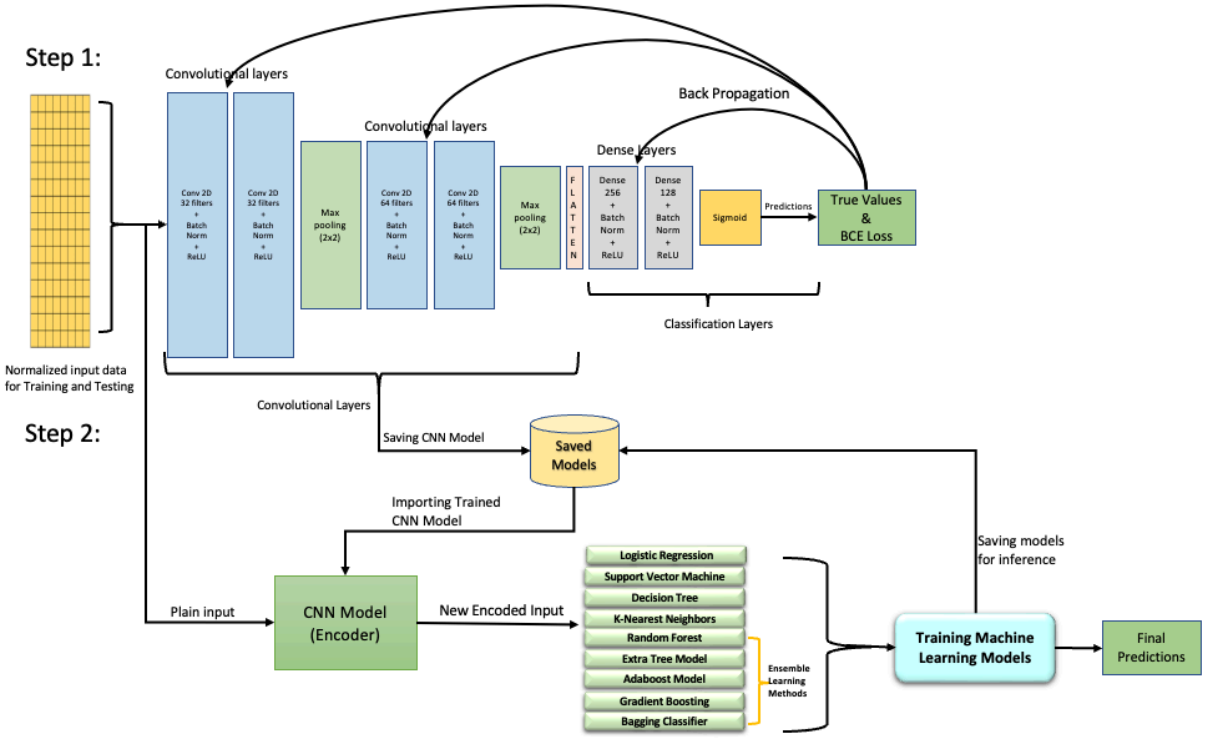

Figure 4.4: Architecture of the Hybrid Model.Step 1 includes training of the convolutional neural networks (CNN) using back propagation.

After the model is trained, the CNN model is stored with the trained weights. In Step 2, the trained CNN model (Encoder) is used to transform the plain input data to new encoded input data. This new input is used to train the machine learning models to generate the final predictions.

5 Performance metrics

This section discusses about the performance metrics that are considered to compare the performances of the models on the test data. These metrics include accuracy, precision, recall, specificity, F1-Score and Receiver Operating Characteristic (ROC) curve.

5.1 Confusion matrix

A confusion matrix is a square matrix which shows the true positive, true negative, false positive and false negative of a classification. In binary classification problems, true positives and true negatives are the instances which are correctly predicted by the algorithm as positive and negative, respectively. In contrast, false positives are instances which are predicted as true by the algorithm but are actually false, and the false negatives are instances which are predicted as false but are actually true.

True Labels

		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Various performance metrics can be calculated using the confusion matrix. These metrics include accuracy, precision, recall, specificity, and F1-score. The accuracy metric gives the percentage of correct predictions made by the model in relation to the total number of predictions made.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{True\ Positive\ (TP) + True\ Negatives\ (TN)}{Total\ number\ of\ samples}$$

Precision is the ratio of the true positive predictions made by the algorithm out of all the positive predictions. A high precision value tells us that the algorithm has generated very few false positives.

$$\text{Precision} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)}$$

Recall is the ratio of the true positive predictions to the total correct cases. Correct cases include the true positive and also the false negative predictions. Recall is also known as sensitivity.

$$\text{Recall} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

Specificity is a measurement which tells us how many of the negative predictions are correctly made. Specificity is the ratio of the true negative predictions to the total number of negative instances in the data set.

$$\text{Specificity} = \frac{True\ Negative\ (TN)}{True\ Negatives\ (TN) + False\ Positives(FP)}$$

The F1-Score is the harmonic mean of the precision and recall values; it ranges from 0 and 1 where a higher F1-scores means that the model has good precision and recall values. If the F1 score is close to 0, the model is performing poorly with imbalance in the precision and the recall values.

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5.2 Receiver Operating Characteristic curve

Receiver Operating Characteristic (ROC) curve is a performance measurement curve for the classification problems. The ROC curve is plotted as True positive rate (TPR; y-axis) vs. False positive rate (FPR; x-axis), where FPR is 1 - specificity and TPR is also known as recall/sensitivity. The points on the ROC curve are plotted at different threshold levels based on the probabilities of the prediction. The Area Under the Curve (AUC) can also be calculated for the ROC curve and indicates how well the model predicted true values as true and false values as false.

6 Results

Several ML GRN prediction models with distinct architectures were assessed, including four supervised learning models (Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors) and five ensemble learning models (Random Forests, Extremely Randomized Trees, AdaBoost Models, Gradient Boosting, Bagging Classifier). Additionally, neural network approaches including FCNs and CNNs with various loss functions and architectures were examined. Hybrid models, which combine machine learning and CNN models, were also implemented, and evaluated. The data employed was compendium transcriptomic data gathered and pooled from public repositories, and the training data sets consisted of known regulatory gene pairs (a regulator and a target gene) and their transcript values extracted from the compendium transcriptomic data, as detailed in Section 3: Multiple OMICS Data Collection.

The following section presents the results of the training procedure and the evaluation of the various models using distinct test data sets. Initially, the outcomes of hyperparameter tuning for ML models are discussed, followed by an examination of cross-validation scores and an examination of the accuracy of ML models on holdout test data, which was 20% of the training data that was retained for validation or testing. These models were trained separately for different species and evaluated using their respective holdout test data sets.

Moreover, a brief overview of the training and tuning for FCNs, CNNs, and hybrid models is provided, including their accuracies on holdout test data, which was 10-20% of the training data that was retained for validation or testing. In addition to using holdout test data, the models were also assessed using real test data sets: *Arabidopsis* Transcriptomic

Test Data Set 1, *Arabidopsis* Transcriptomic Test Data Set 2, the poplar Transcriptomic Test Data Set, and the maize Transcriptomic Test Data Set

6.1 Hyperparameter Tuning and Testing Machine Learning Models

In this study, all ML models were implemented using the scikit-learn library (Pedregosa et al. 2011) in the Python programming language. Hyperparameter tuning is a crucial step in constructing effective ML models since each model has specific adjustable parameters that depend on the data type and distribution. The grid search technique was employed to select hyperparameters by testing various combinations on the models. In this study, *Arabidopsis* training data was used for hyperparameter tuning with the grid search technique. The Table 6.1 presents the various hyperparameters learned for different ML models using the *Arabidopsis* training data.

Table 6.1: Learned hyperparameters and description for different models using *Arabidopsis* training data

Model	**Parameter tuned**	**Description**
Logistic Regression	penalty: l2, C: 1, solver: saga, max iterations: 500	Penalty is a regularization parameter, C is the inverse of regularization strength, solver is the optimization algorithm, max iterations define the number of iterations the algorithm performs on data
Support Vector Machine	C: 0.5, kernel: linear, gamma: scale, degree: 4	C is the regularization parameter, kernel function is used to map data to higher dimensions, gamma controls the shape of the decision boundary, degree defines degree of the polynomial
Decision Tree	criterion: gini, min samples leaf: 5, min samples split: 5	Criterion is used to measure the quality of the split, min samples required to be a leaf node and min samples required to split a node in tree
K-Nearest Neighbors	metric: Manhattan, n neighbors: 6	metric is used to calculate the distance between datapoints, n neighbors define the number of neighbors considered for prediction

Random Forest	maximum depth: 20, min samples leaf: 5, min samples split: 2, estimators: 50	maximum depth is the depth of each Decision Tree, min samples required to be a leaf node and min samples required to split a node in tree, estimators is the number of Decision Trees constructed
Extremely Randomized Trees	maximum depth: 20, min samples leaf: 1, min samples split: 2, estimators: 50	maximum depth is the depth of each Decision Tree, min samples required to be a leaf node and min samples required to split a node in tree, estimators is the number of Decision Trees constructed
AdaBoost	estimators: 100, learning rate: 0.1	estimators value is the number of Decision Trees constructed, learning rate controls the contribution of the weak classifier in final ensemble
Gradient Boosting	estimators: 100, learning rate: 0.1	estimators value is the number of Decision Trees constructed, learning rate controls the contribution of the weak classifier in final ensemble
Bagging Classifier	estimators: 100, max samples: 10, max features: 1	estimators value is the number of Decision Trees constructed, max samples define the maximum samples to use for each copy, max features define the maximum features to use for each copy of the model

The tuned hyperparameters were applied to the ML models for further training and prediction.

After tuning the hyperparameters, 10-fold cross validation was applied to the ML models to compare the performance of the models in different species: *Arabidopsis,* poplar, maize. Figure 6.1 displays the boxplots of the accuracies of the Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine also referred as Support Vector Classifier (SVC), Decision Tree and several ensemble models including Random Forest, Extremely Randomized Trees, the AdaBoost Model, the Gradient Boosting Model, and the Bagging Model. Figure 6.1 A-C presents the accuracies of these models using *Arabidopsis*, poplar, and maize training data, respectively.

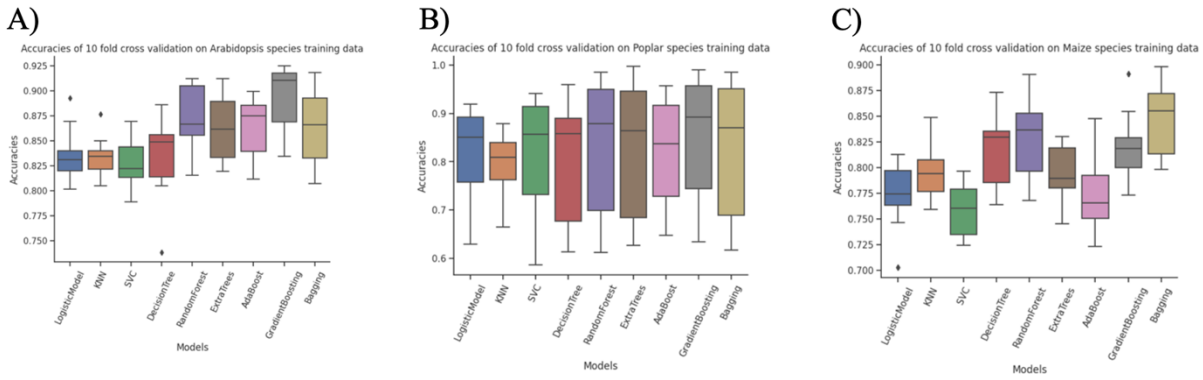

Figure 6.1: Boxplot depicting the accuracies of 10-fold cross-validation on machine learning models. These models include Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, and ensemble models like Random Forest, Extremely Randomized Trees, AdaBoost, Gradient Boosting, and Bagging. A. Accuracies for *Arabidopsis* training data; B. Accuracies for poplar training data; C. Accuracies for maize training data.

The boxplots in Figure 6.1 show the first quartile, median, and third quartile of the accuracy values. Based on the median accuracy score, the ensemble models Random Forests, Gradient Boosting, and Bagging performed better than the other methods.

Next, ML models were trained on 80% of Arabidopsis, poplar, and maize data, reserving 20% as the holdout test set. Table 6.2 displays the accuracies of these models for each species.

Table 6.2: Accuracies of different machine learning models on holdout test data. The models encompass Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), and several ensemble models like Random Forest, Extremely Randomized Trees, AdaBoost, Gradient Boosting, and Bagging. These models were assessed on a holdout test set, comprising 20% of the *Arabidopsis*, poplar, and maize training data that was not utilized during the training phase.

Species	**LR**	**SVM**	**Decision Tree**	**KNN**	**Random Forest classifier**	**Extra Tree classifier**	**AdaBoost Classifier**	**Gradient Boosting**	**Bagging Classifier**
Arabidopsis	80.32	78.29	84.58	82.96	90.67	89.86	90.67	90.67	80.32
Poplar	79.82	77.03	89.72	88.96	96.32	97.72	95.56	93.78	79.82

Maize	76.79	78.61	90.5	83.43	97.37	94.53	84.26	84.88	78.6
Average Scores	78.98	77.98	88.27	85.12	94.79	94.04	90.16	89.78	79.58

At approximately 78%, the linear Logistic Regression and Support Vector Machine models had lower average accuracies with gene expression data in plant species than other models. While the K-Nearest Neighbor Model was more accurate than the linear models, ensemble techniques (Random Forest, Extremely Randomized Trees, AdaBoost classifier, and Gradient Boosting) outperformed all the other techniques on the gene expression data. AdaBoost and Gradient Boosting had average accuracies of approximately 90%. Random Forest and Extremely Randomized Trees had the best performance with an average accuracy of more than 94% on the holdout test data.

6.2 Hyperparameter Tuning and Testing Neural Networks

In this study, neural networks were implemented using the Keras library from TensorFlow (Abadi et al. 2016) in the Python programming language. The construction of a customized FCN for gene expression data requires the selection of multiple hidden layers, each with its own set of hyperparameters. Key hyperparameters to take into account are the learning rate, neuron count, activation function, and batch size. Research has demonstrated that employing multiple layers with a dropout structure in FCN can enhance performance (Chen et al. 2016). In this study, neural networks consisting of two dense layers with 256 and 128 neurons were utilized, each followed by a 10% dropout in the hidden layers of the FCN. An optimized learning rate of 0.00003 was determined through experimentation and

applied to the FCN, CNN, and Hybrid models. Training was conducted using a batch size of 100 and the RMSprop optimizer.

The FCNs were trained using 80% of the training data of *Arabidopsis*, poplar, and maize with different loss functions, including BCE, hinge loss, MSE, MSLE, MAE, Poisson loss, Huber loss, and LogCosh loss, and were tested separately on the 20% holdout test data. This 20% data was not used in the training process. The results are shown in Table 6.3.

Table 6.3: Accuracies of fully connected networks (FCN) with different loss functions. Loss functions include binary cross entropy (BCE), hinge loss, mean squared error (MSE), mean squared logarithmic error (MSLE), mean absolute error (MAE), Poisson loss, Huber loss, and LogCosh loss which were evaluated on the holdout test set. This test set consists of 20% of the *Arabidopsis*, poplar, and maize training data, which was not used during the training phase.

Species	**FCN BCE**	**FCN HINGE**	**FCN MSE**	**FCN MSLE**	**FCN MAE**	**FCN POISSON**	**FCN HUBER**	**FCN LOGCOSH**
Arabidopsis	87.42	84.58	87.02	85.4	89.25	87.62	87.42	87.42
Poplar	95.28	91.85	91.11	92.44	92.92	92.8	92.25	91.37
Maize	89.05	88.82	89.79	89.5	89.05	90.8	90.95	90.71
Average Scores	90.58	88.42	89.31	89.11	90.41	90.41	90.21	89.83

The FCNs with the binary cross entropy (BCE) loss had an average accuracy of 90.58%, a good performance on holdout test data of *Arabidopsis,* poplar, and maize training data. The other loss functions showed approximately the same performance as BCE, with mean absolute error (MAE), Poisson loss, and Huber loss being the closest.

The subsequent step involved conducting hyperparameter tuning on the CNN to identify optimal parameters for training the CNN on the gene expression data. It is essential

to choose the appropriate number of layers and kernels to construct a customized CNN model. To assess which kernel combinations were suitable for the gene expression data, two stacked convolutional layers were used and different numbers of kernels (8, 16, 32, 64, 128, 256) were sequentially applied with a kernel size of 3x3. BCE was utilized as the loss function to train and evaluate these models. Figure 6.2 displays the heatmaps of the accuracy values. As a result, multiple kernel numbers were tested for the three species under investigation.

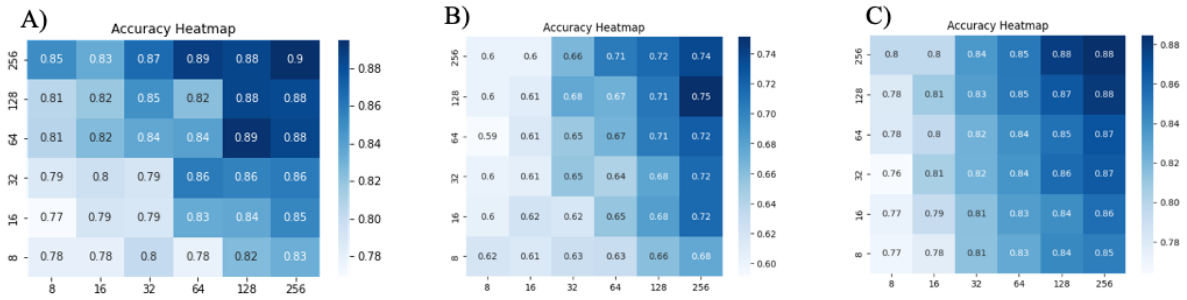

Figure 6.2: Heatmaps of accuracy values of convolutional neural networks (CNN) with different numbers of kernels in the first and second layers of the CNN. The y-axis dimension is the number of kernels in the first layer of the CNN while the x-axis is the number of kernels in the second layer of the CNN. Binary cross-entropy was used as the loss function for the CNN and evaluated on 20% (also referred to as holdout test data) of the *Arabidopsis*, poplar, and maize training data. The scale bars represent a series of accuracy values. A. the accuracy values of the CNN on *Arabidopsis* training data. B. the accuracy values of the CNN on poplar training data. C. the accuracy values of the CNN on maize training data. Note that these 20% holdout test data sets were not used in the training process.

Figure 6.2 reveals that CNN performance with BCE loss improved on *Arabidopsis* data when increasing kernel numbers in first and second layers. However, this enhancement was not seen in poplar and maize species, with 256 kernels in *Arabidopsis* yielding 90% accuracy, while only 74% in poplar. Thus, a higher kernel count doesn't

guarantee increased accuracy for all species. Previous research has indicated that stacking multiple convolutional layers followed by max pooling layers can improve gene expression data prediction performance (Lyudmyla Yasinska-Damri 2022). In this study, multiple convolutional layers were stacked along with max pooling layers, as depicted in Figure 4.3. Further experiments were conducted to alter the number of neurons in the dense layers and using different dropout structures. A dropout value of 0.1 with 256 neurons followed by 128 neurons in the dense layers was ultimately chosen. Various optimizers and learning rate configurations were also explored. The final CNN model employed the RMSprop optimizer with a learning rate of 0.00003.

The CNNs were trained on the 80% training data of *Arabidopsis*, poplar, and maize with different loss functions. In addition to custom-built CNNs, deep convolutional neural networks such as ResNet and MobileNet were also used to predict gene regulatory networks. Table 6.4 shows the accuracies of the CNNs on the 20% holdout test data of the three species.

Table 6.4: Accuracies of Convolutional Neural Networks (CNN) on the holdout test data. Multiple loss functions such as binary cross entropy (BCE), hinge loss, mean squared error (MSE), mean squared logarithmic error (MSLE), mean absolute error (MAE), Poisson loss, Huber loss, and LogCosh loss were utilized in the custom built CNNs. Deep CNNs such as ResNet 50 and MobileNet were also evaluated on the holdout test data. Note that these 20% holdout test data sets in the three species were not used in the training process.

Species	**CNN BCE**	**CNN HINGE**	**CNN MSLE**	**CNN MSE**	**CNN MAE**	**CNN POISSON**	**CNN HUBER**	**CNN LOGCOSH**	**ResNet 50**	**Mobile Net**
Arabidopsis	93.5	91.48	92.29	91.48	91.88	92.69	91.47	92.08	81.93	74.03
Poplar	97.59	98.1	97.59	97.21	98.35	97.85	97.72	96.32	88.67	85.47
Maize	94.86	95.34	90.48	94.08	95.4	88.51	95.65	94.9	85.89	83.07

Average Scores	95.32	94.97	93.45	94.26	95.21	93.02	94.95	94.43	85.5	80.86

Based on the accuracy results for the holdout test data presented in Table 6.4, the CNN model with the BCE loss function was 95.32% accurate and outperformed other models. Other loss functions, such as MAE, Hinge Loss, and Huber loss, displayed similar performance, with average accuracy values of 95.21%, 94.97% and 94.95%, respectively, across the three species. Deep CNNs, including ResNet and MobileNet, did not perform as well as custom-built CNNs with fewer convolutional layers. One potential reason is that these deep neural networks with skip connections require vast amounts of training data, which was not available for any of the species we evaluated. Each of our training data sets was relatively small.

6.3 Training and Testing Hybrid Models

The hybrid architecture training occurs in two stages: training the convolutional encoder model and training the ML model using the output from the convolutional encoder. The convolutional encoder models were trained separately on *Arabidopsis*, poplar, and maize data sets for 100 epochs using the BCE loss function. Figures 6.3 and 6.4 display the accuracy and loss curves of the CNNs. Figure 6.3A-C represents the accuracies of *Arabidopsis*, poplar, maize training data, respectively. Figure 6.4A-C denotes the loss values for the *Arabidopsis*, poplar and maize training data.

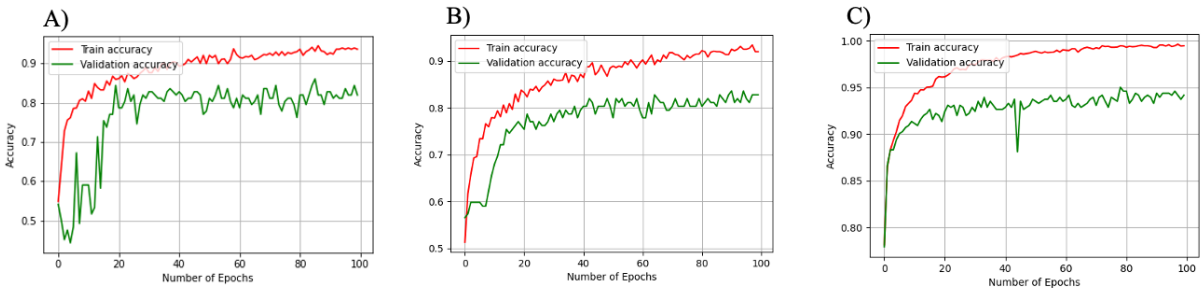

Figure 6.3: Training and validation accuracy curves for CNNs in step 1 of building the hybrid architecture. A. Accuracy of the *Arabidopsis* training data; B. Accuracy of the poplar training data; C. Accuracy of the maize training data. Training accuracy was based on 80% of the overall data, and validation accuracy on 20% of the overall data for *Arabidopsis* (A), poplar (Wilson et al.), and maize (Robertson et al.) training data. The 20% validation data was not used in the training process.

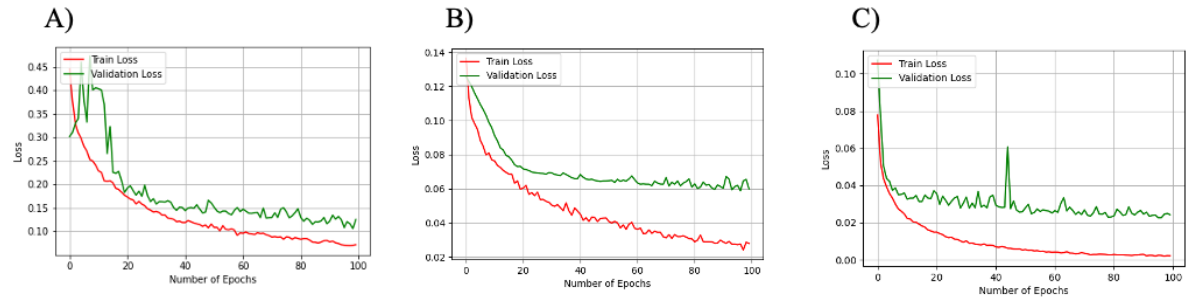

Figure 6.4: Training and validation loss curves for CNNs in the step 1 of building the hybrid architecture. A. Loss values of the *Arabidopsis* training data; B. Loss values of the poplar training data; C. Loss values of the maize training data. Training loss was based on 80% of the overall data, and validation loss on 20% of the overall data for *Arabidopsis* (A), poplar (B) and maize (C) training data. The 20% validation data was not used in the training process.

In comparing the training and validation curves for each of the three species, it is evident that there is no significant gap between them in terms of accuracy, suggesting that the models do not suffer from overfitting. Similarly, the loss curves also followed a consistent pattern. These observations indicate that the models were neither overfitting nor underfitting the data but were effectively generalizing to new data.

The hybrid models were trained on *Arabidopsis*, poplar, and maize species training data. Table 6.5 presents the accuracies of the ML models on the holdout test set based on each species. The holdout test set consists of 20% of the original training data, which was not used for training the models.

Table 6.5: Accuracies of various hybrid models on holdout test data. These models are based on the hybrid architecture and include Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), and ensemble models such as Random Forest, Extremely Randomized Trees, AdaBoost, Gradient Boosting, and Bagging. They were evaluated using one holdout test set, each containing 20% of the *Arabidopsis*, poplar, or maize training data that was not used during the training process.

Species	**LR**	**SVM**	**Decision Tree**	**KNN**	**Random Forest classifier**	**Extra Tree classifier**	**AdaBoost Classifier**	**Gradient Boosting**	**Bagging Classifier**
Arabidopsis	90.67	86.61	85.8	85.19	93.1	91.28	90.65	91.07	90.86
Poplar	97.47	95.94	92.01	88.45	97.97	97.72	98.1	95.94	95.18
Maize	95.31	86.45	95.08	87.73	98.87	97.25	99.13	95.4	86.51
Average Scores	94.48	89.67	90.96	87.12	96.65	95.42	95.96	94.14	90.85

Table 6.5 presents the accuracy results for the hybrid architecture models on the holdout test data. Of the ensemble techniques, Random Forest, Extremely Randomized Trees, and AdaBoost Models exceled, achieving average accuracy scores of 96.65%, 95.42%, and 95.96%, respectively, in the three species. Furthermore, the Logistic Regression Model from the hybrid architecture demonstrates strong performance compared to other methods such as Support Vector Machine, Decision Tree, K-Nearest Neighbors, and Bagging classifier, based on the average accuracies of the three species.

The holdout test results reveal that the Random Forest Model was the top-performer in both regular and hybrid approaches, followed by Extremely Randomized Trees and

AdaBoost models in the hybrid architecture. In the neural networks, BCE loss exhibited better performance on the holdout test data compared to the other loss functions. To further evaluate these models, real test data consisting of the genes involved in the LBP were utilized.

6.4 *In-silico* validation of selected methods

The top-performing hybrid architecture methods on the hold-out validation datasets include Random Forest, Extremely Randomized Trees, and AdaBoost Models. These models utilize input from the CNN encoder in the hybrid architecture. To evaluate the hybrid architecture's performance against standard ML models, Random Forest, Extremely Randomized Trees, and AdaBoost Models which were trained directly on the input without any encoding. This comparison allows for an assessment of hybrid architecture performance relative to conventional ML methods. The models were trained on 80% training data sets for *Arabidopsis*, poplar, and maize separately. Subsequently, the trained models were tested on Transcriptomic Test Data Sets.

Spearman's rank correlation coefficient is often used as a benchmark method in addition to supervised ML models. This statistical measure quantifies the strength of the relationship between two variables and is frequently employed in gene expression analysis to identify associations between genes (Kumari et al. 2012). In this study, Spearman’s rank correlation coefficient was used as a benchmark for lignin pathway analysis. To minimize false discoveries in large data sets, the Benjamini-Hochberg procedure (Benjamini and

Hochberg 1995) was employed to apply False Discovery Rate (FDR) correction after calculating the correlation coefficient.

6.4.1 Test results with *Arabidopsis* Transcriptomic Test Data Set 1

Arabidopsis Transcriptomic Test Data Set 1 consists of genes associated with the lignin biosynthesis pathway (LBP), which plays a role in plant growth and development through the production of lignin, a major component of the cell wall. As detailed in Section 3: Multiple OMICS Data Collection, this transcriptomic data set consists of 28,300 regulatory pairs and the expression data was extracted from the Compendium Data Set 1 with 1,253 samples.

The Hybrid Random Forest and plain Random Forest Models were both fine-tuned and trained using *Arabidopsis* training data. In the hybrid model, a CNN encoder processes the training data and passes it to the ML model, while the plain ML model utilizes unencoded data. Transcriptomic Test Data Set 1 serves as the testing data for LBP analysis in *Arabidopsis*. Probability values for the positive regulation class were extracted and sorted in descending order, leading to the selection of the top 1000 positively regulated gene pairs (TFs and target genes) from the Transcriptomic Test Data Set 1. The unique TFs within the top 1000 pairs were tallied, with each TF assigned a frequency value. This value was then used to identify the top 50 TFs that positively regulate target genes. Similar training and testing experiments were conducted using Hybrid Extremely Randomized Trees, plain Extremely Randomized Trees, Hybrid AdaBoost Model, and plain AdaBoost Model. Table 6.6 presents the top 50 TFs for both the Hybrid Random Forest and plain Random Forest Models.

Table 6.6: Comparison of the top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by the Hybrid Random Forest and the Plain Random Forest Models on *Arabidopsis* Transcriptomic Test Data Set 1. The frequency of each TF within the top 1000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid Random Forest Model			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	AT3G08500_MYB83	20	(Zhong and Ye, 2012)
2	AT1G71930_VND7	20	(Yamaguchi et al., 2011)
3	AT4G36920_AP2	20	-
4	AT2G20180_AtbHLH15	20	-
5	AT5G11260_HY5	20	-
6	AT5G16560_KAN	20	-
7	AT1G24260_SEPALLATA3	20	-
8	AT1G32770_SND1/NST3	19	(Mitsuda et al., 2007)
9	AT1G14350_FLP	19	-
10	AT5G12870_MYB46	18	(Zhong and Ye, 2012)
11	AT2G02820_AtMYB88	18	-
12	AT4G23810_AtWRKY53	16	-
13	AT3G27920_GL1	14	-
14	AT5G62380_VND6	10	(Ohashi-Ito, Oda, and Fukuda, 2010)
15	AT1G24625_ZFP7	10	-
16	AT1G74930_ORA47	9	-
17	AT2G43010_AtbHLH9	8	-
18	AT5G13790_AGL15	7	-
19	AT3G24650_ABI3	6	-
20	AT4G18960_AG	6	-
21	AT3G02310_AGL4	6	-
22	AT1G69120_AP1	6	-
23	AT1G26310_CAL	6	-
24	AT2G44730	6	-
25	AT3G54340_AP3	5	-
26	AT1G69180_CRC	5	-
27	AT5G10120_EIL4	5	-
28	AT1G23420_INO	5	-
29	AT1G01060_LHY	5	-
30	AT5G57520_ZFP2	5	-
31	AT2G40220_ABI4	4	-
32	AT5G15800_AGL2	4	-
33	AT2G45650_AGL6	4	-
34	AT2G16910_AMS	4	-
35	AT1G25340_AtMYB116	4	-
36	AT1G12610_DDF1	4	-
37	AT1G47870_E2FC	4	-
38	AT3G13960_GRF5	4	-
39	AT2G33880_HB3	4	-
40	AT5G62020_HSF6	4	-
41	AT1G67100_LOB40	4	-
42	AT2G46770_NST1	4	(Mitsuda et al., 2007)
43	AT5G20240_PI	4	-
44	AT4G27330_SPL	4	-
45	AT2G44745_WRKY12	4	(Wang et al., 2010)
46	AT1G10480_ZFP5	4	-
47	AT2G45420_LBD18	3	-
48	AT3G13890_MYB26	3	(Yang et al., 2007)
49	AT5G57620_MYB36	3	(Kamiya et al., 2015)
50	AT2G18060_VND1	3	(Zhou, Zhong, and Ye, 2014)

Plain Random Forest Model			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	AT4G36920_AP2	20	-
2	AT5G16560_KAN	20	-
3	AT2G20180_AtbHLH15	20	-
4	AT5G11260_HY5	20	-
5	AT2G44730	20	-
6	AT1G71930_VND7	17	(Yamaguchi et al., 2011)
7	AT3G08500_MYB83	17	(Zhong and Ye, 2012)
8	AT5G12870_MYB46	16	(Zhong and Ye, 2012)
9	AT1G66140_ZFP4	14	-
10	AT1G24260_SEPALLATA3	12	-
11	AT3G13890_MYB26	12	(Caiyun Yang, 2007)
12	AT2G02820_AtMYB88	11	-
13	AT1G14350_FLP	11	-
14	AT1G32770_SND1/NST3	10	(Mitsuda et al., 2007)
15	AT5G17300	8	-
16	AT2G32370	6	-
17	AT1G25340_AtMYB116	6	-
18	AT1G25330_AtbHLH75	6	-
19	AT4G27330	6	-
20	AT2G18060	6	-
21	AT2G40220_ABI4	6	-
22	AT1G23420_INO	6	-
23	AT4G35700	6	-
24	AT5G18450	6	-
25	AT2G44745	6	-
26	AT4G00220_LBD30	6	-
27	AT3G27920_GL1	6	-
28	AT1G09540_MYB61	6	-
29	AT2G44745_WRKY12	6	(Wang et al., 2010)
30	AT4G00220	6	-
31	AT2G18060_VND1	6	(Zhou, Zhong, and Ye, 2014)
32	AT1G09540	6	-
33	AT1G12610	6	-
34	AT5G62380_VND6	6	(Ohashi-Ito, Oda, and Fukuda, 2010)
35	AT3G06120_AtbHLH45	6	-
36	AT4G09960_STK	5	-
37	AT3G30530	5	-
38	AT3G01530	5	-
39	AT1G61110	5	-
40	AT2G42830_SHP2	5	-
41	AT5G03790_LMI1	5	-
42	AT1G15360_SHINE1	5	-
43	AT1G66380_AtMYB114	5	-
44	AT1G35490	5	-
45	AT5G23260_AGL32	5	-
46	AT2G46770_NST1	5	(Mitsuda et al., 2007)
47	AT5G15800_AGL2	5	-
48	AT1G26310_CAULIFLOWER	5	-
49	AT4G18960_AG	5	-
50	AT5G53210_AtbHLH98	5	-

Table 6.6 provides the frequency values of each TF within the top 1000 predictions for both the hybrid and plain Random Forest Models. TFs highlighted in red are identified as true regulators in the current literature with relevant references provided. AT3G08500_MYB83 and AT5G12870_MYB46 are recognized as master regulators in secondary cell wall (SCW) biosynthesis in *Arabidopsis thaliana* (Zhong and Ye 2012). Another study identified AT2G46770_NST1 and AT1G32770_SND1/NST3 as key regulators in SCW biosynthesis (Mitsuda et al. 2007). The Hybrid Random Forest Model

successfully detected these crucial TFs involved in LBP. Additionally, it identified AT2G18060_VND1 (Zhou, Zhong, and Ye 2014), AT5G62380_VND6 (Ohashi-Ito, Oda, and Fukuda 2010), AT1G71930_VND7 (Yamaguchi et al. 2011), AT3G13890_MYB26 (Yang et al. 2007), AT5G57620_MYB36 (Kamiya et al. 2015) and AT2G44745_WRKY12 (Wang et al. 2010), all of which are known to be authentic regulators in *Arabidopsis* thaliana. The Hybrid Random Forest Model detected 10 real TFs, while the plain Random Forest Model identified nine true positives. This demonstrates that the Hybrid Random Forest Model outperformed the plain Random Forest Model.

The Extremely Randomized Tree Model was also applied in both hybrid and plain architectures. The results are presented in Table 6.7.

Table 6.7: Comparison of the top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by the Hybrid Extremely Randomized Trees and the Plain Extremely Randomized Trees Models on *Arabidopsis* Transcriptomic Test Data Set 1. The frequency of each TF within the top 1000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid Extremely Randomized Trees Model			
Rank	Transcription Factor	Freq.	Reference
1	AT3G08500_MYB83	20	(Zhong and Ye, 2012)
2	AT1G71930_VND7	20	(Yamaguchi et al., 2011)
3	AT4G36920_AP2	20	-
4	AT5G05410_DREB2A	20	-
5	AT2G20180_AtbHLH15	20	-
6	AT5G16560_KAN	20	-
7	AT1G14350_FLP	20	-
8	AT1G24625_ZFP7	20	-
9	AT2G02820_AtMYB88	20	-
10	AT4G23810_AtWRKY53	20	-
11	AT5G11260_HY5	20	-
12	AT1G32770_SND1/NST3	19	(Mitsuda et al., 2007)
13	AT1G24260_SEPALLATA3	19	-
14	AT2G44730	17	-
15	AT1G74930	17	-
16	AT5G12870_MYB46	15	(Zhong and Ye, 2012)
17	AT1G21910	13	-
18	AT5G62380_VND6	11	-
19	AT1G47870_E2FC	10	(Ohashi-Ito, Oda, and Fukuda, 2010)
20	AT2G43010_AtbHLH9	10	-
21	AT4G23980_ARF9	9	-
22	AT1G12610	8	-
23	AT3G59060_AtbHLH65	8	-
24	AT5G13790_AGL15	7	-
25	AT1G01060_LHY	7	-
26	AT4G37260_AtMYB73	7	(Manoj Kumar, 2015)
27	AT5G20240_PI	6	-
28	AT2G40470_LBD15	6	-
29	AT2G40470	6	-
30	AT2G46830_CCA1	6	-
31	AT3G54340_AP3	6	-
32	AT5G54680	6	-
33	AT4G22680	6	-
34	AT4G22680_MYB85	6	(Geng et al., 2020)
35	AT2G16910	5	-
36	AT4G18960_AG	5	-
37	AT5G62020_HSF6	5	-
38	AT1G61730	5	-
39	AT5G10510_AIL6	5	-
40	AT1G69120_AP1	5	-
41	AT1G77450	5	-
42	AT2G22840_AtGRF1	5	-
43	AT3G02310_AGL4	5	-
44	AT5G17800_AtMYB56	5	-
45	AT4G12350_MYB42	4	(Geng et al., 2020)
46	AT1G09540	4	-
47	AT2G18060_VND1	4	(Zhou, Zhong, and Ye, 2014)
48	AT2G18060	4	-
49	AT2G45420	4	-
50	AT4G00220_LBD30	4	-

Plain Extremely Randomized Trees Model			
Rank	Transcription Factor	Freq.	Reference
1	AT4G36920_AP2	20	-
2	AT2G02820_AtMYB88	20	-
3	AT1G14350_FLP	20	-
4	AT1G24625_ZFP7	20	-
5	AT5G11260_HY5	20	-
6	AT5G16560_KAN	20	-
7	AT2G20180_AtbHLH15	20	-
8	AT1G74930	18	-
9	AT5G05410_DREB2A	18	-
10	AT2G44730	15	-
11	AT1G71930_VND7	14	(Yamaguchi et al., 2011)
12	AT1G24260_SEPALLATA3	14	-
13	AT5G12870_MYB46	13	(Zhong and Ye, 2012)
14	AT3G08500_MYB83	13	(Zhong and Ye, 2012)
15	AT2G40470	12	-
16	AT2G40470_LBD15	12	-
17	AT3G51910_HSFA7A	12	-
18	AT2G45420_LBD18	12	-
19	AT2G18060_VND1	12	-
20	AT1G21910	12	-
21	AT2G45420	12	-
22	AT2G18060	12	-
23	AT4G23810_AtWRKY53	12	-
24	AT1G47870_E2FC	11	-
25	AT5G62380_VND6	10	(Ohashi-Ito, Oda, and Fukuda, 2010)
26	AT1G12260	10	-
27	AT1G12260_VND4	10	(Zhou, Zhong, and Ye, 2014)
28	AT1G12610	10	-
29	AT1G32770_SND1/ NST3	9	(Mitsuda et al., 2007)
30	AT3G61850_DAG1 BBFa	9	-
31	AT1G09540_MYB61	8	-
32	AT1G09540	8	-
33	AT2G22840_AtGRF1	8	-
34	AT5G15840	7	-
35	AT3G27010_TCP20	7	-
36	AT5G54680	7	-
37	AT1G01060_LHY	7	-
38	AT5G10510_AIL6	7	-
39	AT5G62020_HSF6	7	-
40	AT4G17490_AtERF6	7	-
41	AT1G77450_At1g77450	6	-
42	AT2G46830_CCA1	6	-
43	AT3G06120_AtbHLH45	6	-
44	AT5G13790_AGL15	6	-
45	AT3G02310_AGL4	6	-
46	AT4G00220	6	-
47	AT2G47890_COL13	6	-
48	AT4G00220_LBD30	6	-
49	AT2G46770_NST1	5	(Mitsuda et al., 2007)
50	AT5G25830_GATA-12	5	-

Table 6.7 displays the frequency values of each TF within the top 1000 predictions for both the hybrid and plain Extremely Randomized Tree Models. TFs highlighted in red are acknowledged as true regulators in the current literature, with corresponding references provided. In addition to AT3G08500_MYB83, AT5G12870_MYB46, and AT1G32770_SND1/NST3, which are known master regulators in SCW biosynthesis, the Hybrid Extra Tree Model identified AT4G37260_AtMYB73 (Kumar, Campbell, and Turner 2016), AT4G22680_MYB85 (Geng et al. 2020), and AT4G12350_MYB42 (Geng et al. 2020), which were not detected by the Random Forest Models. Similar to the Hybrid Random Forest Model, the Hybrid Extra Tree Model detected AT2G18060_VND1 (Zhou, Zhong, and Ye 2014), AT5G62380_VND6 (Ohashi-Ito, Oda, and Fukuda 2010), and

AT1G71930_VND7 (Yamaguchi et al. 2011). It is noteworthy that the plain Extra Tree Model identified VND4 (Zhou, Zhong, and Ye 2014), which was not detected by other models. In total, the Hybrid Extremely Randomized Tree Model identified nine TFs, whereas the plain model detected only seven. Consequently, the hybrid model performed better than the plain model.

The AdaBoost model was also applied to both hybrid and plain architectures, with results presented in Table 6.8.

Table 6.8: Comparison of the top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by the Hybrid AdaBoost and the Plain AdaBoost Models on *Arabidopsis* Transcriptomic Test Data Set 1. The frequency of each TF within the top 1000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid AdaBoost Model			
Rank	Transcription Factor	Freq.	Reference
1	AT1G32770_SND1/NST3	20	(Mitsuda et al., 2007)
2	AT3G08500_MYB83	20	(Zhong and Ye, 2012)
3	AT1G71930_VND7	20	(Yamaguchi et al., 2011)
4	AT5G12870_MYB46	20	(Zhong and Ye, 2012)
5	AT4G36920_AP2	20	-
6	AT5G11260_HY5	20	-
7	AT5G16560_KAN	20	-
8	AT1G24260_SEPALLATA3	20	-
9	AT1G14350_FLP	20	-
10	AT2G02820_AtMYB88	20	-
11	AT4G23810_AtWRKY53	20	-
12	AT2G20180_AtbHLH15	20	-
13	AT1G74930	19	-
14	AT1G01060_LHY	18	-
15	AT3G27920_GL1	16	-
16	AT1G24625_ZFP7	16	-
17	AT3G02310_AGL4	12	-
18	AT2G44730	12	-
19	AT3G59060_AtbHLH65	10	-
20	AT1G69120_AP1	10	-
21	AT4G18960_AG	10	-
22	AT5G20240_PI	8	-
23	AT3G51910_HSFA7A	8	-
24	AT5G62380_VND6	8	(Ohashi-Ito, Oda, and Fukuda, 2010)
25	AT2G46830_CCA1	7	-
26	AT2G43010_AtbHLH9	7	-
27	AT5G13790_AGL15	6	-
28	AT4G22680	6	-
29	AT2G45420	6	-
30	AT4G22680_MYB85	6	(Geng et al., 2020)
31	AT3G54340_AP3	5	-
32	AT2G16910	5	-
33	AT2G40470_LBD15	4	-
34	AT4G00870_AtbHLH14	4	-
35	AT1G25330_AtbHLH75	4	-
36	AT3G13960	4	-
37	AT5G05410_DREB2A	4	-
38	AT4G00220_LBD30	4	-
39	AT2G40220_ABI4	4	-
40	AT2G40470	4	-
41	AT4G00220	4	-
42	AT2G33880	4	-
43	AT1G09540	4	-
44	AT2G18060	4	-
45	AT3G58190	4	-
46	AT1G09540_MYB61	4	-
47	AT1G15360_SHINE1	4	-
48	AT5G15800_AGL2	4	-
49	AT2G45420_LBD18	4	-
50	AT2G18060_VND1	4	(Zhou, Zhong, and Ye, 2014)

Plain AdaBoost Model			
Rank	Transcription Factor	Freq.	Reference
1	AT1G56650_MYB75	38	-
2	AT4G36920_AP2	19	-
3	AT5G16560_KAN	16	-
4	AT2G20180_AtbHLH15	15	-
5	AT5G11260_HY5	14	-
6	AT4G25990	13	-
7	AT2G46680_ATHB-7	12	-
8	AT5G62470_AtMYB96	11	-
9	AT5G43840_HSFA6A	11	-
10	AT2G41940_ZFP8	9	-
11	AT3G61890_Athb-12	8	-
12	AT1G66390_AtMYB90	8	-
13	AT5G56620_anac099	7	-
14	AT2G02820_AtMYB88	7	-
15	AT3G27920_GL1	7	-
16	AT2G36080	7	-
17	AT1G05230	7	-
18	AT5G56860_GNC	7	-
19	AT3G08500_MYB83	7	(Zhong and Ye, 2012)
20	AT3G24140	7	-
21	AT1G71930_VND7	7	(Yamaguchi et al., 2011)
22	AT5G17300	6	-
23	AT1G14440	6	-
24	AT4G22680_MYB85	6	(Geng et al., 2020)
25	AT4G22680	6	-
26	AT1G32770_SND1/NST3	5	(Mitsuda et al., 2007)
27	AT5G46830	5	-
28	AT5G65320	5	-
29	AT1G02340_AtbHLH26	5	-
30	AT1G75520	5	-
31	AT4G32280_IAA29	5	-
32	AT1G52890_ANAC019	5	-
33	AT5G53420	5	-
34	AT2G39250_SNZ	5	-
35	AT2G26580_YABBY5	5	-
36	AT4G31615	5	-
37	AT1G14350_FLP	5	-
38	AT3G58190	5	-
39	AT3G58070	5	-
40	AT1G18750	5	-
41	AT5G26930_GATA-23	5	-
42	AT1G13960_AtWRKY4	5	-
43	AT3G02380_COL2	5	-
44	AT1G75240_Athb-33	5	-
45	AT5G67180	5	-
46	AT1G74500	5	-
47	AT1G77920	5	-
48	AT3G15500_AtNAC3	5	-
49	AT5G62430	5	-
50	AT1G71030_ATMYBL2	5	-

Table 6.8 displays the frequency value of each TF within the top 1000 predictions for both Hybrid and Plain AdaBoost Models. TFs highlighted in red are recognized as true regulators based on current literature, with corresponding references provided. Like other supervised models, the Hybrid AdaBoost Model successfully detected AT3G08500_MYB83, AT5G12870_MYB46, AT1G32770_SND1/NST3, AT1G71930_VND7, AT5G62380_VND6, AT4G22680_MYB85, and AT2G18060_VND1 as true regulators. In contrast, the Plain AdaBoost Model identified only three TFs: AT3G08500_MYB83, AT1G32770_SND1/NST3, AT1G71930_VND7, and AT4G22680_MYB85. The Hybrid AdaBoost Model not only predicted more TFs than

the plain model, but also ranked the TFs higher with greater frequency. As a result, the hybrid model outperformed the traditional AdaBoost model.

To compare the performance of the models with a baseline model, Spearman's rank correlation coefficient is commonly used as a benchmark method alongside supervised ML models. This statistical measure assesses the strength of the relationship between two variables and is frequently used in gene expression analysis to identify associations between genes. To reduce false discoveries in large datasets, the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) was applied to implement a False Discovery Rate (FDR) correction after calculating the correlation coefficient. Gene regulatory pairs were sorted in ascending order based on the corrected p-value, ranging from the most to least significant relationship, and the top 1000 regulatory pairs were selected from the list. The unique TFs within the top 1000 pairs were counted, and each TF was assigned a frequency value. This value was then used to identify the top 50 TFs that positively regulate target genes. Table 6.9 displays the top 50 TFs as well as their corresponding ranks.

Table 6.9: Top 50 transcription factors (TFs) that regulate lignin biosynthesis pathway based on the corrected Spearman correlation coefficient on *Arabidopsis* Transcriptomic Test Data Set 1. The frequency of each TFs within the top 1000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Spearman Correlation Coefficient			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	AT5G60100	6	-
2	AT4G13640	6	-
3	AT3G50700	6	-
4	AT1G64530	6	-
5	AT1G20693	6	-
6	AT1G04250	6	-
7	AT5G37020_ARF8	5	-
8	AT4G31060	5	-
9	AT3G58680_ATMBF1B	5	-
10	AT3G23210	5	-
11	AT3G21175	5	-
12	AT2G34710_AtHB-14	5	-
13	AT2G01650	5	-
14	AT1G71692	5	-
15	AT1G67970	5	-
16	AT1G49720_ABF1	5	-
17	AT1G19270	5	-
18	AT5G63280	4	-
19	AT5G53200_TRY	4	-
20	AT5G46910	4	-
21	AT5G41920_AtGRAS-28	4	-
22	AT5G13080	4	-
23	AT4G34610	4	-
24	AT4G17900	4	-
25	AT4G00050_AtbHLH16	4	-
26	AT3G54620	4	-
27	AT3G17609	4	-
28	AT3G16280	4	-
29	AT3G02830_ZFN1	4	-
30	AT2G43000	4	-
31	AT2G40740_AtWRKY55	4	-
32	AT2G37630_AtMYB91	4	-
33	AT2G16720_MYB7	4	(Wang et al., 2010)
34	AT1G70000	4	-
35	AT1G22070	4	-
36	AT1G17460_TRFL3	4	-
37	AT1G12260_VND4	4	Zhou, Zhong, and Ye, 2014
38	AT1G04550	4	-
39	AT5G67480	3	-
40	AT5G66630	3	-
41	AT5G65410_Athb-25	3	-
42	AT5G63080	3	-
43	AT5G61380_TOC1	3	-
44	AT5G60120	3	-
45	AT5G58010	3	-
46	AT5G57620	3	-
47	AT5G56860_GNC	3	-
48	AT5G54230	3	-
49	AT4G00180_YAB3	3	-
50	AT1G10200_WLIM1	3	-

From Table 6.9, it was evident that the Spearman correlation coefficient method identified only two TFs, namely AT2G16720_MYB7 and AT1G12260_VND4. Hybrid models have not only outperformed traditional ML models but have also demonstrated significantly better performance than the benchmark Spearman correlation coefficient method, a proven, highly efficacious method for identifying pathway regulators (Kumari et al. 2012).

Finally, a GRN was built using the best-performing hybrid ML model, the Hybrid Random Forest model, on the Arabidopsis Transcriptomic Test Data 1. In the Figure 6.5, the target genes of the LBP are represented by green nodes, while the top 50 TFs with the

highest connectivity are depicted by the other nodes based on the frequency count. The light coral nodes specifically denote the true transcription factors involved in the LBP.

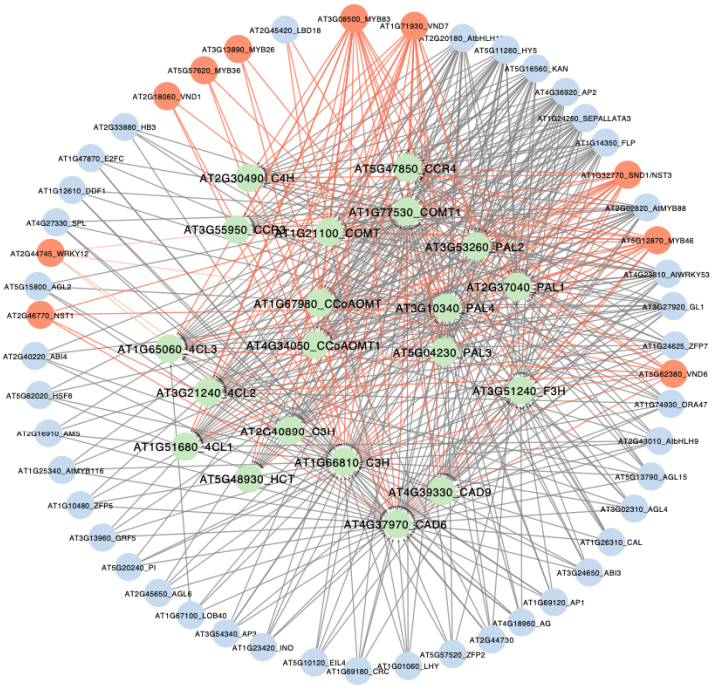

Figure 6.5: Regulatory network generated by the Hybrid Random Forest model on *Arabidopsis* Transcriptomic Test Data Set 1. The green nodes denote the target genes, all the other nodes are the top 50 transcription factors based on the frequency count. The light coral nodes represent the true TFs which involve in the lignin biosynthesis pathway.

6.4.2 Test results with *Arabidopsis* Transcriptomic Test Data Set 2

Arabidopsis Transcriptomic Test Data Set 2 was adopted from Taylor-Teeples's Supplementary Table 2 (Taylor-Teeples et al. 2015). As detailed in Section 3: Multiple OMICS Data Collection, this transcriptomic data set consists of 1,164 regulatory pairs out of which 582 are considered to be positive regulatory pairs as they were validated by using Yeast One Hybrid System (Bulyk et al. 1999) and the remaining 582 regulatory pairs are generated by random sampling of the TFs with other targets which are not shown as positive in the AGRIS (Yilmaz et al. 2011). The expression data was extracted from the

Compendium Data Set 1 which has 1,253 samples. Random Forest, Extremely Randomized Trees, and AdaBoost Models from both hybrid architecture and conventional ML methods were employed for the predictions. The Table 6.10 shows the accuracy, precision, recall, specificity, f1-score, area under curve (AUC) score.

Table 6.10: Accuracy, precision, recall, specificity, F1-score, and Area under curve (AUC) score for *Arabidopsis* Transcriptomic Test Data Set 2. The data set contains 1,164 regulatory pairs, with 582 positive regulatory pairs and 582 negative regulatory pairs.

No.	**Model**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**	**AUC score**
1	Random Forest Classifier Hybrid	83.26	83.33	83.26	85.59	83.25	93.00
2	Random Forest Classifier Plain	84.55	86.19	84.54	95.20	84.37	89.80
3	Extra Trees Classifier Hybrid	85.15	85.38	85.15	89.19	85.12	93.31
4	Extra Trees Classifier Plain	84.03	85.60	84.03	94.51	83.85	88.05
5	AdaBoost Classifier Hybrid	81.20	81.33	81.20	84.39	81.18	91.84
6	AdaBoost Classifier Plain	84.98	85.23	84.97	89.19	84.95	89.38

The results in the Table 6.10 reveal that tree-based models, including Random Forest and Extremely Randomized Trees, demonstrate strong performance in the hybrid architecture with area under the curve (AUC) scores of 93.00% and 93.31%, respectively. Moreover, hybrid models consistently outperformed their plain ML counterparts. Figure 6.6 illustrates the ROC curve for Random Forest, Extremely Randomized Trees (Extra Trees), and AdaBoost, using both hybrid architecture and plain methods.

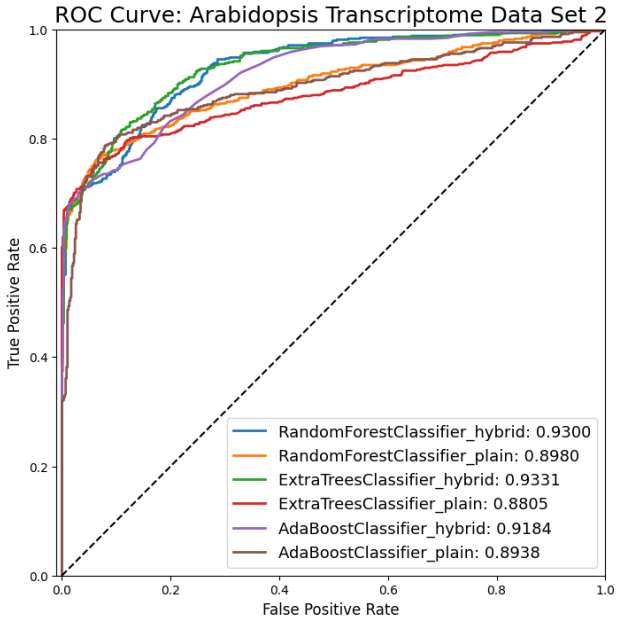

Figure 6.6: The Receiver Operating Characteristic (ROC) curves using multiple models on the *Arabidopsis* Transcriptomic Test Data Set 2. The figure displays the predictions for Random Forest, Extremely Randomized Trees (Extra Trees), and AdaBoost using both hybrid architecture and plain methods. The true positive rate and false positive rate were calculated based on the 1,164 regulatory pairs found in *Arabidopsis* Transcriptomic Test Data Set 2.

The ROC curve was drawn using the true positive rate and the false positive rate and primarily used to compare different models. The predictions made by the Hybrid Random Forest Model, based on Taylor-Teeples's Supplementary Table 2 data (Taylor-Teeples et al. 2015) are shown in Figure 6.7. Out of 582 positive regulatory pairs, the Hybrid Random Forest Model identified 471 pairs with true relationships. In contrast, the Hybrid Extremely Randomized Trees and Hybrid AdaBoost Models detected 443 and 425 pairs, respectively.

The Plain Random Forest Model predicted 420 positive regulatory pairs, while the plain Extremely Randomized Trees and plain AdaBoost Models identified 435 and 464 regulatory pairs, respectively. As a result, the Hybrid Random Forest Model outperformed other methods and achieved a higher AUC ROC score. Although the Extra Tree Model and

Random Forest Model from the hybrid architecture displayed similar AUC scores, the Random Forest Model detected true relationships more accurately. The heatmap in Figure 6.7 demonstrates the Hybrid Random Forest Model probabilities ranging from 0.5 to 1. Columns correspond to various TFs, and rows represent different target genes.

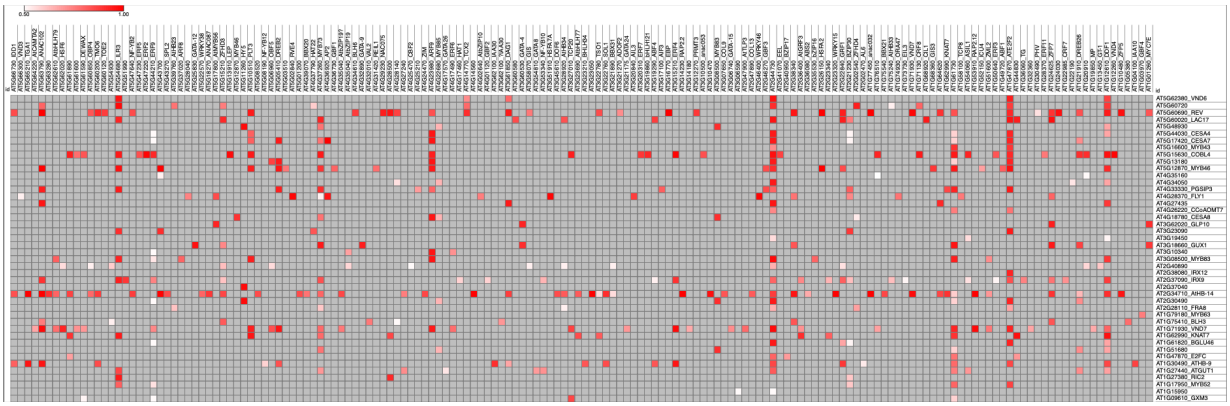

Figure 6.7: Heatmap of the prediction probabilities of the Hybrid Random Forest Model on Arabidopsis Transcriptomic Test Data Set 2. In the heatmap, only the predictions of positive regulatory pairs with probability values greater than 0.5 were considered. The data set comprises 1,164 regulatory pairs, with 582 positive regulatory pairs and 582 negative regulatory pairs. Columns represent various transcription factors, while rows indicate different target genes.

Upon examining Transcriptomic Test Data Sets 1 and 2, it was evident that the hybrid models consistently outperformed their traditional counterparts in terms of accuracy and predictive power. The Hybrid Random Forest, Extremely Randomized Trees, and AdaBoost Models demonstrated superior results when compared to traditional ML models. For example, in *Arabidopsis* Transcriptomic Test Data Set 1, the Hybrid Random Forest Model detected 10 true TFs, whereas the plain Random Forest Model identified only 9. Additionally, the Hybrid Extremely Randomized Trees Model identified 9 TFs, while the

plain model detected only seven. Similarly, the Hybrid AdaBoost Model detected seven true TFs, whereas the plain model identified only four true positives.

In *Arabidopsis* Transcriptomic Test Data Set 2, the Hybrid Random Forest Model achieved an AUC ROC score of 93% and detected 471 out of 582 positive regulatory pairs, whereas the plain architecture model only predicted 420 positive pairs. Similarly, the Hybrid Extremely Randomized Trees Model achieved an AUC ROC score of 93.31% and identified 443 pairs, compared to the plain model's detection of 435 pairs. The Hybrid AdaBoost Model detected 425 pairs with an AUC ROC score of 91.84%, while the plain model identified 464 pairs with an AUC ROC score of 89.38%.

The enhanced performance of hybrid models demonstrates their potential for more accurate and reliable predictions in gene regulatory network analysis in the *Arabidopsis* species. To further test the methods, poplar Transcriptomic Test Data Set and maize Transcriptomic Test Data Set was used

6.4.3 Test results with Poplar Transcriptomic Test Data Set

The poplar Transcriptomic Test Data Set focuses on genes related to the LBP. Through homologous mapping of *Arabidopsis* lignin pathway target genes to poplar, a total of 25 target genes for poplar species were identified. These 25 target genes were paired with 1,717 unique TFs, resulting in a total of 42,925 regulatory pairs in the poplar Transcriptomic Test Data Set. The gene expression data for the poplar Transcriptomic Test Data Set, consisting of 743 samples, were extracted from the NCBI SRA database (Compendium Data Set 2).

Following the approach used for *Arabidopsis*, both hybrid and plain Random Forest Models were fine-tuned and trained on poplar training data. The hybrid model uses encoded data from the CNN encoder for training ML models, while the plain model uses unencoded data to train ML models. The poplar Transcriptomic Test Data Set serves as the testing data. The top 2000 positive regulatory gene pairs were selected, and their unique TFs were tallied. This frequency value helps to identify the top 50 TFs positively regulating target genes. Similar training and testing experiments were conducted using hybrid and plain Extremely Randomized Trees as well as hybrid and plain AdaBoost Models. The top 50 TFs of the Hybrid Random Forest Model and plain Random Forest Model are shown in Table 6.11.

Table 6.11: Comparison of the Top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by Hybrid Random Forest and Plain Random Forest Models on poplar Transcriptomic Test Data Set. The frequency of each TF within the top 2000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid Random Forest Model			
Rank	Transcription Factor	Freq.	Reference
1	Potri.009G061500_MYB83	24	(Zhong and Ye, 2012)
2	Potri.019G083600_VND7	24	(Yamaguchi et al., 2011)
3	Potri.001G267300_MYB83	24	(Zhong and Ye, 2012)
4	Potri.013G113100_VND7	24	(Yamaguchi et al., 2011)
5	Potri.002G252800_AtbHLH15	24	-
6	Potri.002G023400_E2FC	24	-
7	Potri.018G029500_HY5	24	-
8	Potri.001G088600_ARF9	24	-
9	Potri.013G153400_ATWRKY33	24	-
10	Potri.011G007800_AtWRKY42	24	-
11	Potri.006G105300_ATWRKY33	24	-
12	Potri.002G139500_At2g44730	24	-
13	Potri.010G093000_FLP	24	-
14	Potri.009G055700_AGL15	24	-
15	Potri.008G148400_FLP	24	-
16	Potri.004G064300_AG	24	-
17	Potri.002G164400_AtWRKY22	24	-
18	Potri.001G058400_SEPALLATA3	24	-
19	Potri.001G092900_AtWRKY53	24	-
20	Potri.012G031700_AtWRKY53	24	-
21	Potri.016G128300_ATWRKY33	24	-
22	Potri.014G090300_AtWRKY22	24	-
23	Potri.002G168700_AtWRKY53	24	-
24	Potri.006G251800_HY5	24	-
25	Potri.005G140700_AP2	24	-
26	Potri.007G046200_AP2	24	-
27	Potri.017G137600_KAN	24	-
28	Potri.003G138600_AtWRKY53	24	-
29	Potri.004G082400_KAN	24	-
30	Potri.003G169600_SEPALLATA3	24	-
31	Potri.011G075800_AG	24	-
32	Potri.014G096200_AtWRKY53	24	-
33	Potri.019G123500_ATWRKY33	23	-
34	Potri.001G154200_AtERF6	22	-
35	Potri.004G007500_AtWRKY42	22	-
36	Potri.001G327100_TCP20	21	-
37	Potri.002G172800_ARF9	21	-
38	Potri.017G068748_TCP20	21	-
39	Potri.016G083600_TTG2	21	-
40	Potri.014G051200_At2g44730	20	-
41	Potri.003G167900_TCP20	19	-
42	Potri.001G060000_TCP20	19	-
43	Potri.017G082900_Athb-34	19	-
44	Potri.003G080600_AtERF6	18	-
45	Potri.001G080900_At2g43000	18	-
46	Potri.015G075600_GL1	18	-
47	Potri.002G180800_LHY	17	-
48	Potri.001G044500_AtWRKY40	17	-
49	Potri.002G172101_AtbHLH13	16	-
50	Potri.002G055400_AtbHLH65	15	-

Plain Random Forest Model			
Rank	Transcription Factor	Freq.	Reference
1	Potri.002G252800_AtbHLH15	24	-
2	Potri.005G140700_AP2	24	-
3	Potri.006G251800_HY5	24	-
4	Potri.017G137600_KAN	24	-
5	Potri.007G046200_AP2	24	-
6	Potri.004G082400_KAN	24	-
7	Potri.002G172800_ARF9	24	-
8	Potri.018G029500_HY5	24	-
9	Potri.001G267300_MYB83	24	(Zhong and Ye, 2012)
10	Potri.014G096200_AtWRKY53	24	-
11	Potri.002G168700_AtWRKY53	24	-
12	Potri.003G138600_AtWRKY53	23	-
13	Potri.008G210900	20	-
14	Potri.008G148400_FLP	19	-
15	Potri.019G083600_VND7	18	(Yamaguchi et al., 2011)
16	Potri.013G113100_VND7	18	(Yamaguchi et al., 2011)
17	Potri.006G105300_ATWRKY33	17	-
18	Potri.013G153400_ATWRKY33	16	-
19	Potri.003G167900_TCP20	14	-
20	Potri.001G327100_TCP20	14	-
21	Potri.019G123500_ATWRKY33	13	-
22	Potri.002G164400_AtWRKY22	13	-
23	Potri.001G060000_TCP20	12	-
24	Potri.009G061500_MYB83	12	(Zhong and Ye, 2012)
25	Potri.017G068748_TCP20	11	-
26	Potri.006G221800_TT2	11	-
27	Potri.001G092900_AtWRKY53	10	-
28	Potri.012G031700_AtWRKY53	9	-
29	Potri.010G093000_FLP	9	-
30	Potri.004G007500_AtWRKY42	8	-
31	Potri.014G017300_AZF3	8	-
32	Potri.001G154200_AtERF6	8	-
33	Potri.012G138900_HSF6	7	-
34	Potri.006G263600_WRKY60	7	-
35	Potri.003G182200_AtWRKY40	7	-
36	Potri.001G079800_At5g51190	7	-
37	Potri.002G139500_At2g44730	7	-
38	Potri.015G075600_GL1	6	-
39	Potri.002G023400_E2FC	6	-
40	Potri.001G088600_ARF9	6	-
41	Potri.003G142100_ARF9	6	-
42	Potri.003G165000_At5g13220	6	-
43	Potri.016G083600_TTG2	6	-
44	Potri.002G119300_AZF3	6	-
45	Potri.016G128300_ATWRKY33	5	-
46	Potri.003G151000_At5g61590	5	-
47	Potri.009G141600	5	-
48	Potri.009G055700_AGL15	5	-
49	Potri.014G090300_AtWRKY22	5	-
50	Potri.011G075800_AG	4	-

Table 6.11 presents the frequency values of each TF within the top 2000 predictions for both hybrid and plain Random Forest Models. TFs highlighted in red were recognized as true regulators in the current literature, with pertinent references provided via homologous mapping. The Hybrid Random Forest Model effectively detected key TFs involved in lignin biosynthesis, such as Potri.009G061500_MYB83 and Potri.001G267300_MYB83, which are known as master regulators in SCW biosynthesis (Zhong and Ye 2012). Additionally, the model identified Potri.019G083600_VND7 and Potri.013G113100_VND7, which are also involved in the LBP (Yamaguchi et al. 2011). While both hybrid and plain Random Forest Models detected the same number of TFs, the hybrid model assigned higher rankings to positive TFs. In addition to these two models,

Extremely Randomized Trees were employed in both hybrid and plain architectures. The top 50 TFs were extracted in a similar manner using frequency values. The results are shown in Table 6.12.

Table 6.12: Comparison of the Top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by the Hybrid Extremely Randomized Trees and the Plain Extremely Randomized Trees Models on poplar Transcriptomic Test Data Set. The frequency of each TF within the top 2000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid Extremely Randomized Trees Model			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	Potri.007G046200_AP2	24	-
2	Potri.002G168700_AtWRKY53	24	-
3	Potri.001G088600_ARF9	24	-
4	Potri.002G139500_At2g44730	24	-
5	Potri.003G169600_SEPALLATA3	24	-
6	Potri.002G172800_ARF9	24	-
7	Potri.011G075800_AG	24	-
8	Potri.001G092900_AtWRKY53	24	-
9	Potri.005G140700_AP2	24	-
10	Potri.003G138600_AtWRKY53	24	-
11	Potri.017G137600_KAN	24	-
12	Potri.014G096200_AtWRKY53	24	-
13	Potri.004G082400_KAN	24	-
14	Potri.006G251800_HY5	24	-
15	Potri.002G252800_AtbHLH15	24	-
16	Potri.018G029500_HY5	24	-
17	Potri.001G058400_SEPALLATA3	24	-
18	Potri.001G267300_MYB83	23	(Zhong and Ye, 2012)
19	Potri.008G148400_FLP	23	-
20	Potri.009G055700_AGL15	23	-
21	Potri.002G164400_AtWRKY22	23	-
22	Potri.004G064300_AG	23	-
23	Potri.014G051200_At2g44730	22	-
24	Potri.013G153400_ATWRKY33	22	-
25	Potri.014G090300_AtWRKY22	22	-
26	Potri.002G023400_E2FC	21	-
27	Potri.017G082900_Athb-34	21	-
28	Potri.009G061500_MYB83	20	(Zhong and Ye, 2012)
29	Potri.010G093000_FLP	20	-
30	Potri.003G167900_TCP20	20	-
31	Potri.003G142100_ARF9	19	-
32	Potri.012G031700_AtWRKY53	18	-
33	Potri.011G132400	18	-
34	Potri.006G105300_ATWRKY33	18	-
35	Potri.004G007500_AtWRKY42	18	-
36	Potri.016G128300_ATWRKY33	17	-
37	Potri.013G113100_VND7	17	(Yamaguchi et al., 2011)
38	Potri.014G100100_ARF9	17	-
39	Potri.011G007800_AtWRKY42	16	-
40	Potri.008G113200	16	-
41	Potri.017G016700_SND2	16	(Hussey et al., 2011)
42	Potri.006G221800_TT2	15	-
43	Potri.019G083600_VND7	15	(Yamaguchi et al., 2011)
44	Potri.001G080900_At2g43000	14	-
45	Potri.015G075600_GL1	14	-
46	Potri.002G172101_AtbHLH13	14	-
47	Potri.014G066100_At3g60580	14	-
48	Potri.001G154200_AtERF6	14	-
49	Potri.007G135300_SND2	14	(Hussey et al., 2011)
50	Potri.003G080600_AtERF6	13	-

Plain Extremely Randomized Trees Model			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	Potri.002G252800_AtbHLH15	24	-
2	Potri.006G251800_HY5	24	-
3	Potri.005G140700_AP2	24	-
4	Potri.018G029500_HY5	23	-
5	Potri.008G210900	23	-
6	Potri.007G046200_AP2	22	-
7	Potri.017G137600_KAN	21	-
8	Potri.004G082400_KAN	21	-
9	Potri.002G172800_ARF9	21	-
10	Potri.002G168700_AtWRKY53	20	-
11	Potri.001G267300_MYB83	19	(Zhong and Ye, 2012)
12	Potri.017G107500	19	-
13	Potri.001G327100_TCP20	19	-
14	Potri.019G123500_ATWRKY33	18	-
15	Potri.004G108320	16	-
16	Potri.014G096200_AtWRKY53	16	-
17	Potri.008G148400_FLP	16	-
18	Potri.003G138600_AtWRKY53	15	-
19	Potri.006G221800_TT2	15	-
20	Potri.002G023400_E2FC	15	-
21	Potri.009G061500_MYB83	15	(Zhong and Ye, 2012)
22	Potri.013G153400_ATWRKY33	15	-
23	Potri.014G100100_ARF9	15	-
24	Potri.010G093000_FLP	14	-
25	Potri.002G139500_At2g44730	14	-
26	Potri.011G007800_AtWRKY42	14	-
27	Potri.014G090300_AtWRKY22	14	-
28	Potri.003G167900_TCP20	14	-
29	Potri.002G164400_AtWRKY22	13	-
30	Potri.012G031700_AtWRKY53	13	-
31	Potri.002G180800_LHY	13	-
32	Potri.014G051200_At2g44730	12	-
33	Potri.001G088600_ARF9	12	-
34	Potri.001G154200_AtERF6	12	-
35	Potri.017G068748_TCP20	12	-
36	Potri.016G083600_TTG2	12	-
37	Potri.004G007500_AtWRKY42	12	-
38	Potri.006G133200_TTG2	12	-
39	Potri.001G092900_AtWRKY53	12	-
40	Potri.019G083600_VND7	12	(Yamaguchi et al., 2011)
41	Potri.010G006800	12	-
42	Potri.011G132400	11	-
43	Potri.009G055700_AGL15	11	-
44	Potri.013G113100_VND7	11	(Yamaguchi et al., 2011)
45	Potri.002G055400_AtbHLH65	10	-
46	Potri.003G151000_At5g61590	9	-
47	Potri.001G079600_At5g61590	9	-
48	Potri.003G142100_ARF9	9	-
49	Potri.006G005500_At5g47640	9	-
50	Potri.012G138900_HSF6	9	-

The hybrid and plain Extremely Randomized Tree Models detected Potri.009G061500_MYB83, Potri.001G267300_MYB83, Potri.019G083600_VND7, and

Potri.013G113100_VND7, akin to the Random Forest Model. These models also identified SND2 genes, which are involved in SCW formation (Hussey et al. 2011). The Hybrid Extremely Randomized Trees detected both Potri.017G016700_SND2 and Potri.007G135300_SND2. As a result, the hybrid model outperformed the traditional model by identifying six TFs, while the plain model detected only four. In addition, the Hybrid Extremely Randomized Trees Model outperformed the Hybrid Random Forest Model in poplar. This indicates that the additional variance added by the Extremely Randomized Trees Model is useful for gene expression data in poplar species.

Hybrid and plain AdaBoost Models were applied to the poplar Transcriptomic Test Data Set in a similar fashion, and the results are presented in Table 6.13. The hybrid AdaBoost Model was able to detect Potri.009G061500_MYB83 and Potri.019G083600_VND7, which were also detected by the other models. The hybrid AdaBoost Model also identified Potri.003G022800_XND1, an important regulator in the LBP (Zhao et al. 2008); this TF was not identified by the other models in poplar species. In contrast, the plain AdaBoost Model detected only one in the top 50 TFs which is Potri.019G083600_VND7, indicating that using the CNN encoder to encode the gene expression data has clear advantages for inferring gene regulatory relationships.

Table 6.13: Comparison of the top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by the Hybrid AdaBoost Model and the Plain AdaBoost Model on the poplar Transcriptomic Test Data Set. The frequency of each TF within the top 2000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid AdaBoost Model			
Rank	Transcription Factor	Freq.	Reference
1	Potri.007G046200_AP2	24	-
2	Potri.018G029500_HY5	24	-
3	Potri.002G252800_AtbHLH15	24	-
4	Potri.017G137600_KAN	23	-
5	Potri.003G169600_SEPALLATA3	22	-
6	Potri.019G083600_VND7	21	(Yamaguchi et al., 2011)
7	Potri.014G096200_AtWRKY53	21	-
8	Potri.011G075800_AG	20	-
9	Potri.002G023400_E2FC	16	-
10	Potri.014G106800_LHY	14	-
11	Potri.010G093000_FLP	11	-
12	Potri.014G051200_At2g44730	11	-
13	Potri.014G058600_GATA-4	11	-
14	Potri.017G068748_TCP20	11	-
15	Potri.015G075800_GL1	10	-
16	Potri.005G176000_At1g21910	10	-
17	Potri.014G100100_ARF9	10	-
18	Potri.015G141100_HSF6	9	-
19	Potri.010G036400	8	-
20	Potri.003G080600_AtERF6	8	-
21	Potri.004G115500_AGL2	8	-
22	Potri.016G094800_At2g37000	8	-
23	Potri.018G049600_TT2	8	-
24	Potri.008G166200_ERF1	8	-
25	Potri.003G022800_XND1	7	(Zhao et al., 2008)
26	Potri.015G009632_SDG34	7	-
27	Potri.003G022800_At5g64530	7	-
28	Potri.003G000400	7	-
29	Potri.002G157600_AtMYB17	7	-
30	Potri.009G055700_AGL15	7	-
31	Potri.001G159200_At1g31040	7	-
32	Potri.007G066800	7	-
33	Potri.005G149100_At5g65590	7	-
34	Potri.014G094500_At2g46310	7	-
35	Potri.018G044900_GATA-12	7	-
36	Potri.013G081200	7	-
37	Potri.004G082400_KAN	7	-
38	Potri.010G119900	7	-
39	Potri.011G103100	7	-
40	Potri.006G054500	7	-
41	Potri.019G040900	7	-
42	Potri.007G044600_AtbHLH88	7	-
43	Potri.005G118000_AP3	7	-
44	Potri.010G184400	7	-
45	Potri.010G185700	7	-
46	Potri.008G142000	7	-
47	Potri.009G061500_MYB83	6	(Zhong and Ye, 2012)
48	Potri.007G053500_AtGRAS18	6	-
49	Potri.002G149000_LBD18	6	-
50	Potri.002G149000_At2g45420	6	-

Plain AdaBoost Model			
Rank	Transcription Factor	Freq.	Reference
1	Potri.007G046200_AP2	21	-
2	Potri.002G252800_AtbHLH15	19	-
3	Potri.018G029500_HY5	13	-
4	Potri.017G137600_KAN	12	-
5	Potri.003G036900_At1g55110	12	-
6	Potri.001G082700_AtGRF8	9	-
7	Potri.014G096200_AtWRKY53	8	-
8	Potri.014G075200_ANL2	8	-
9	Potri.001G255532_At3g51950	8	-
10	Potri.005G148400_ANT	8	-
11	Potri.003G169600_SEPALLATA3	8	-
12	Potri.008G148200_Athb-13	6	-
13	Potri.018G065400_At3g13960	6	-
14	Potri.014G106800_LHY	6	-
15	Potri.010G093000_FLP	6	-
16	Potri.006G143200_At3g13960	4	-
17	Potri.006G167700	4	-
18	Potri.001G238400_OBP2	4	-
19	Potri.016G121800	4	-
20	Potri.010G099200_At2g02070	4	-
21	Potri.004G082400_KAN	4	-
22	Potri.004G135100	4	-
23	Potri.014G007200_AtGRF2	4	-
24	Potri.006G057200_AtbHLH60	4	-
25	Potri.018G129800_YABBY5	4	-
26	Potri.003G219900	4	-
27	Potri.010G093400_Athb-13	4	-
28	Potri.003G096300_KAN2	4	-
29	Potri.018G008500_At2g24570	4	-
30	Potri.004G116100_TFPD	4	-
31	Potri.003G138600_AtWRKY53	4	-
32	Potri.018G091600	4	-
33	Potri.005G207200_AtbHLH65	4	-
34	Potri.012G104900_AtbHLH137	4	-
35	Potri.005G140700_AP2	4	-
36	Potri.002G035200_Athb-33	4	-
37	Potri.018G054700_ATH1	4	-
38	Potri.014G055700_At2g44940	4	-
39	Potri.019G083600_VND7	3	(Yamaguchi et al., 2011)
40	Potri.001G041400_At1g16070	3	-
41	Potri.008G210900	3	-
42	Potri.012G018300_SDG34	3	-
43	Potri.014G152000_At1g05230	3	-
44	Potri.007G109900	3	-
45	Potri.005G192000_At1g76890	3	-
46	Potri.016G128300_ATWRKY33	3	-
47	Potri.010G167500_AtMYB110	3	-
48	Potri.011G116800_SPL4	3	-
49	Potri.014G107200_At3g61970	3	-
50	Potri.004G060900_WRKY	3	-

To compare the performance of the models with a baseline model, Spearman's rank correlation coefficient was applied on the poplar transcriptomic test data, similar to the *Arabidopsis* species. The Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) was applied to implement a False Discovery Rate (FDR) correction after calculating the correlation coefficient. Gene regulatory pairs were sorted in ascending order based on the corrected p-value, ranging from the most to least significant relationship, and the top 2000 regulatory pairs were selected from the list. The unique TFs within the top 2000 pairs were counted, and each TF was assigned a frequency value. This value was then used to identify

the top 50 TFs that positively regulate target genes. Table 6.14 displays the top 50 TFs as well as their corresponding ranks.

Table 6.14: Top 50 transcription factors (TFs) that regulate lignin biosynthesis pathway based on the corrected Spearman correlation coefficient on poplar Transcriptomic Test Data Set. The frequency of each TFs within the top 2000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Spearman Correlation Coefficient			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	Potri.013G039100_At5g28300	14	-
2	Potri.018G129800_YABBY5	14	-
3	Potri.003G036900_At1g55110	13	-
4	Potri.001G137800_At5g46880	13	-
5	Potri.014G152000_At1g05230	13	-
6	Potri.004G020400	13	-
7	Potri.004G230800	13	-
8	Potri.019G045900	13	-
9	Potri.014G107200_At3g61970	12	-
10	Potri.011G083100	12	-
11	Potri.015G075600_GL1	12	-
12	Potri.014G099900_At1g01250	12	-
13	Potri.014G037200_KAN4	12	-
14	Potri.015G104200_AtbHLH137	12	-
15	Potri.015G022000_TRY	12	-
16	Potri.003G046700	12	-
17	Potri.013G054000_NAC	12	-
18	Potri.003G169100	12	-
19	Potri.012G104900_AtbHLH137	12	-
20	Potri.016G136500	11	-
21	Potri.005G205400_At2g43000	11	-
22	Potri.002G041700	11	-
23	Potri.005G246700	11	-
24	Potri.017G094800_TFPD	11	-
25	Potri.004G050150_ARF3	10	-
26	Potri.010G223300_GATA-8	10	-
27	Potri.007G014400_VND2	10	(Zhou, Zhong, and Ye, 2014)
28	Potri.006G152700	10	-
29	Potri.018G068700	10	-
30	Potri.017G016700_SND2	10	(Hussey et al., 2011)
31	Potri.002G141200_At2g44940	10	-
32	Potri.017G082000_At1g26870	10	-
33	Potri.004G159300_At2g16400	10	-
34	Potri.010G099100_At2g02070	10	-
35	Potri.007G135300_SND2	10	(Hussey et al., 2011)
36	Potri.017G119900_C3H14	10	(Chai et al., 2015)
37	Potri.012G126500	10	-
38	Potri.001G137600_KAN2	10	-
39	Potri.002G181600_At3g61970	10	-
40	Potri.014G066100_At3g60580	10	-
41	Potri.017G139500	10	-
42	Potri.008G106700	10	-
43	Potri.002G154700_ANL2	10	-
44	Potri.007G014400	10	-
45	Potri.001G112200_KNAT7	10	(Yu, 2019)
46	Potri.002G034600	10	-
47	Potri.004G095100_C3H14	9	(Chai et al., 2015)
48	Potri.017G137600_KAN	9	-
49	Potri.005G192000_At1g76890	9	-
50	Potri.014G080900	9	-

In Table 6.14, it is evident that the Spearman metric identified Potri.017G016700_SND2 and Potri.007G135300_SND2, like the Hybrid Extremely Randomized Trees Model. Additionally, it detected Potri.017G119900_C3H14 and Potri.004G095100_C3H14, which play a vital role in the LBP (Chai et al. 2015). It also

identified Potri.001G112200_KNAT7, a regulator of SCW biosynthesis growth (Yu 2019). In poplar species, as in *Arabidopsis*, hybrid models consistently outperformed traditional ML models as seen in Tables 6.11 to 6.13. Although the Spearman metric exhibits comparable performance with the hybrid models, it has some downsides such as being sensitive to outliers and assuming monotonic relationships between variables. These limitations make hybrid models more effective in identifying key TFs in poplar species.

Finally, a GRN was built using the best-performing hybrid ML model, the Hybrid Extremely Randomized Trees model, on the poplar Transcriptomic Test Data. In Figure 6.8, the target genes of the LBP are represented by green nodes, while the top 50 TFs with the highest connectivity are depicted by the other nodes based on the frequency count. The light coral nodes specifically denote the true transcription factors involved in the LBP.

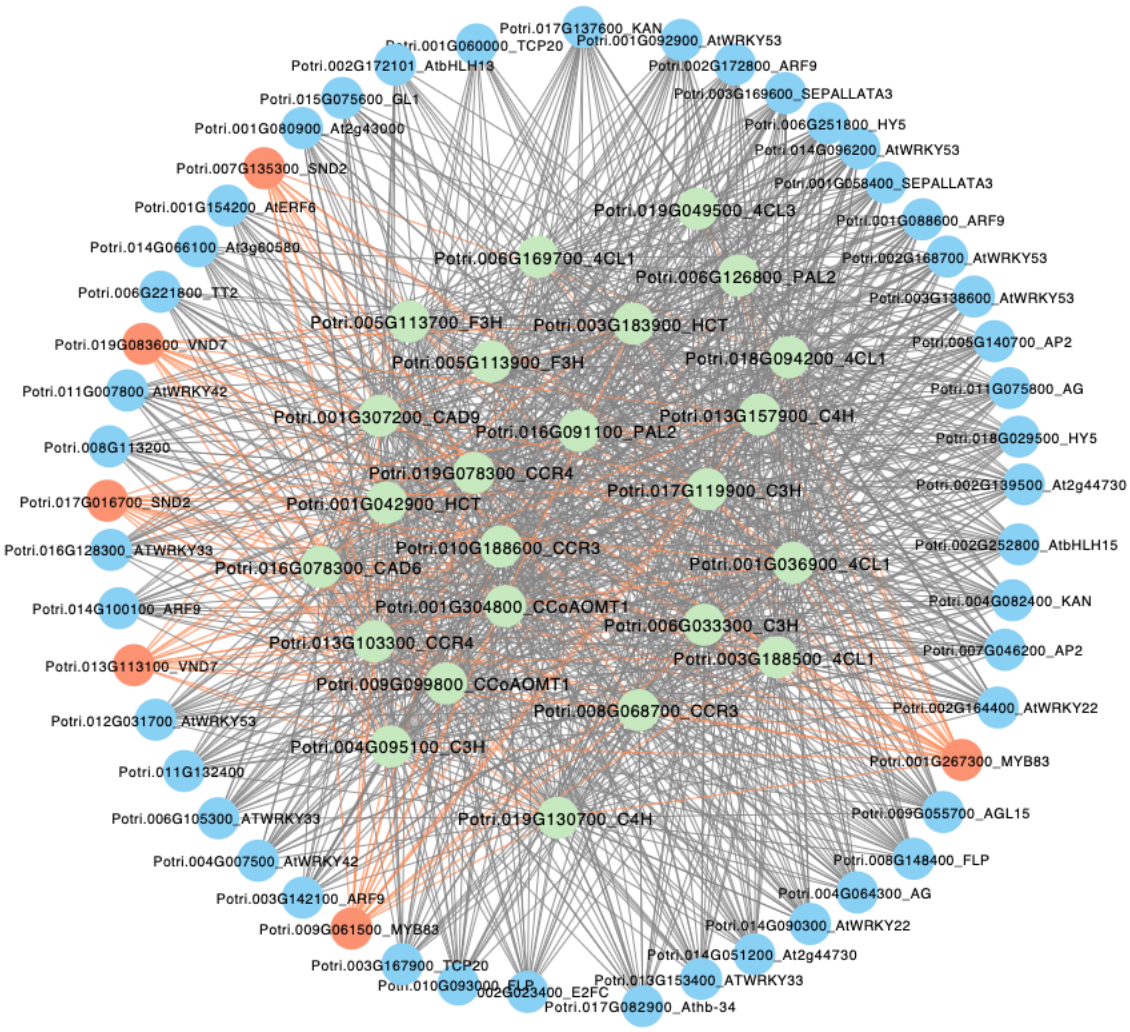

Figure 6.8: Regulatory network generated by the Hybrid Extremely Randomized Trees model on poplar Transcriptomic Test Data Set. The green nodes denote the target genes, all the other nodes are the top 50 transcription factors based on the frequency count. The light coral nodes represent the true TFs which involve in the lignin biosynthesis pathway.

6.4.4 Test results with Maize Transcriptomic Test Data Set

The maize Transcriptomic Test Data Set was compiled with genes related to the LBP. Homologous mapping of *Arabidopsis* lignin pathway target genes to maize resulted in a total of 38 target genes for the maize species. These target genes were paired with 2,555 unique TFs, yielding a total of 97,090 regulatory pairs in the maize Transcriptomic Test

Data Set. Gene expression data from 1,626 samples for the maize Transcriptomic Test Data Set were extracted from the NCBI SRA database (Compendium Data Set 3).

Following the approach used for *Arabidopsis* and poplar species, both hybrid and plain Random Forest Models were fine-tuned and trained on maize training data. The hybrid model employs data encoded by the CNN encoder, while the plain model utilizes unencoded data. The maize Transcriptomic Test Data Set was used for testing. The top 2000 positive regulatory gene pairs were chosen, and their unique TFs were counted. This frequency value assists in identifying the top 50 TFs positively influencing target genes. Comparable training and testing experiments were carried out using hybrid and plain Extremely Randomized Trees Models as well as hybrid and plain AdaBoost Models. Table 6.15 displays the top 50 TFs for the Hybrid and plain Random Forest Models.

Table 6.15: Comparison of the top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by the Hybrid Random Forest and the Plain Random Forest Models on maize Transcriptomic Test Data Set. The frequency of each TF within the top 2000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid Random Forest Model			
Rank	Transcription Factor	Freq.	Reference
1	Zm00001eb176840_VND7	34	(Yamaguchi et al., 2011)
2	Zm00001eb260000_VND7	34	(Yamaguchi et al., 2011)
3	Zm00001eb076470_VND7	34	(Yamaguchi et al., 2011)
4	Zm00001eb122740_HY5	34	-
5	Zm00001eb070520_AtbHLH15	34	-
6	Zm00001eb432100_AP2	34	-
7	Zm00001eb213550_AtbHLH9	34	-
8	Zm00001eb284010_AG	34	-
9	Zm00001eb338060_AG	34	-
10	Zm00001eb102450_SEPALLATA3	34	-
11	Zm00001eb098330_AtWRKY53	34	-
12	Zm00001eb317770_SEPALLATA3	34	-
13	Zm00001eb036590_SEPALLATA3	34	-
14	Zm00001eb004600_AtbHLH15	34	-
15	Zm00001eb400120_SEPALLATA3	34	-
16	Zm00001eb050790_AtbHLH15	34	-
17	Zm00001eb068520_AP2	34	-
18	Zm00001eb265610_AP2	34	-
19	Zm00001eb424050_HY5	34	-
20	Zm00001eb311960_KAN	34	-
21	Zm00001eb424040_HY5	34	-
22	Zm00001eb336820_HY5	34	-
23	Zm00001eb385610_HY5	34	-
24	Zm00001eb062460_AP2	34	-
25	Zm00001eb387280_AP2	34	-
26	Zm00001eb355240_AP2	34	-
27	Zm00001eb235510_HY5	34	-
28	Zm00001eb093920_MYB46	33	(Zhong and Ye, 2012)
29	Zm00001eb145460_AG	33	-
30	Zm00001eb327040_AP1	33	-
31	Zm00001eb312620_Athb-30	33	-
32	Zm00001eb410950_MYB46	32	(Zhong and Ye, 2012)
33	Zm00001eb406030_AtWRKY53	32	-
34	Zm00001eb120710_AG	30	-
35	Zm00001eb184340	28	-
36	Zm00001eb359470_AtWRKY53	27	-
37	Zm00001eb172450_LHY	27	-
38	Zm00001eb118120_AP1	26	-
39	Zm00001eb203940_AtWRKY22	26	-
40	Zm00001eb344160_AtWRKY53	26	-
41	Zm00001eb310270_AtWRKY53	24	-
42	Zm00001eb134890_ZFP7	24	-
43	Zm00001eb109820_FLP	23	-
44	Zm00001eb415760_LHY	22	-
45	Zm00001eb344810_AtWRKY22	22	-
46	Zm00001eb358680_AtWRKY22	21	-
47	Zm00001eb376400_AtWRKY53	21	-
48	Zm00001eb199110_At2g44730	20	-
49	Zm00001eb159410_AtWRKY22	19	-
50	Zm00001eb289170	19	-

Plain Random Forest Model			
Rank	Transcription Factor	Freq.	Reference
1	Zm00001eb122740_HY5	34	-
2	Zm00001eb387280_AP2	34	-
3	Zm00001eb068520_AP2	34	-
4	Zm00001eb312620_Athb-30	34	-
5	Zm00001eb134890_ZFP7	34	-
6	Zm00001eb284010_AG	34	-
7	Zm00001eb317770_SEPALLATA3	34	-
8	Zm00001eb327040_AP1	34	-
9	Zm00001eb036590_SEPALLATA3	34	-
10	Zm00001eb311960_KAN	34	-
11	Zm00001eb102450_SEPALLATA3	34	-
12	Zm00001eb338060_AG	34	-
13	Zm00001eb432100_AP2	34	-
14	Zm00001eb336820_HY5	34	-
15	Zm00001eb265610_AP2	34	-
16	Zm00001eb235510_HY5	34	-
17	Zm00001eb070520_AtbHLH15	34	-
18	Zm00001eb355240_AP2	34	-
19	Zm00001eb385610_HY5	34	-
20	Zm00001eb050790_AtbHLH15	34	-
21	Zm00001eb062460_AP2	34	-
22	Zm00001eb076470_VND7	33	(Yamaguchi et al., 2011)
23	Zm00001eb184340	33	-
24	Zm00001eb213550_AtbHLH9	33	-
25	Zm00001eb424040_HY5	31	-
26	Zm00001eb424050_HY5	31	-
27	Zm00001eb098330_AtWRKY53	29	-
28	Zm00001eb138380_SHP2	28	-
29	Zm00001eb322390_At5g60200	27	-
30	Zm00001eb400120_SEPALLATA3	27	-
31	Zm00001eb142960_TCP20	26	-
32	Zm00001eb310270_AtWRKY53	25	-
33	Zm00001eb411130_SHP2	24	-
34	Zm00001eb406030_AtWRKY53	24	-
35	Zm00001eb145460_AG	24	-
36	Zm00001eb199110_At2g44730	23	-
37	Zm00001eb319350_HSF6	23	-
38	Zm00001eb033570_At2g44730	22	-
39	Zm00001eb430640_AtERF4	22	-
40	Zm00001eb004600_AtbHLH15	21	-
41	Zm00001eb120710_AG	19	-
42	Zm00001eb023870_At2g44730	19	-
43	Zm00001eb109820_FLP	19	-
44	Zm00001eb008690_AGL4	19	-
45	Zm00001eb100610_At1g61730	18	-
46	Zm00001eb359470_AtWRKY53	18	-
47	Zm00001eb344160_AtWRKY53	17	-
48	Zm00001eb415760_LHY	17	-
49	Zm00001eb387370_AtMYB73	15	(Kumar, Campbell, and Turner, 2016)
50	Zm00001eb118120_AP1	15	-

Table 6.15 displays the frequency values of each TF within the top 2000 predictions for both hybrid and plain Random Forest Models. TFs highlighted in red are acknowledged as true regulators in current literature, with relevant references provided through homologous mapping. The Hybrid Random Forest Model effectively identified vital TFs involved in lignin biosynthesis, such as Zm00001eb176840_VND7, Zm00001eb260000_VND7, and Zm00001eb076470_VND7, which play critical roles in the LBP (Yamaguchi et al. 2011). Furthermore, it also detected Zm00001eb093920_MYB46 and Zm00001eb410950_MYB46, known as master regulators in secondary cell wall (SCW) biosynthesis (Zhong and Ye 2012). In contrast, the plain Random Forest Model identified only two true positive TFs,

Zm00001eb076470_VND7 and Zm00001eb387370_AtMYB73. Consequently, the Hybrid Random Forest Model demonstrated significantly better performance than the traditional ML model.

The Extremely Randomized Trees Model was also utilized in both hybrid and plain architectures. The top 50 TFs were extracted in a similar fashion using frequency values. The results are presented in Table 6.16.

Table 6.16: Comparison of the top 50 transcription factors (TFs) predicted that regulate lignin biosynthesis pathway by the Hybrid Extremely Randomized Trees and the Plain Extremely Randomized Trees Models on maize Transcriptomic Test Data Set. The frequency of each TFs within the top 2000 predicted regulatory relationships was counted to represents how many pathway genes it might have influenced. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid Extremely Randomized Trees Model			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	Zm00001eb355240_AP2	34	-
2	Zm00001eb235510_HY5	34	-
3	Zm00001eb317770_SEPALLATA3	34	-
4	Zm00001eb036590_SEPALLATA3	34	-
5	Zm00001eb122740_HY5	34	-
6	Zm00001eb387280_AP2	34	-
7	Zm00001eb432100_AP2	34	-
8	Zm00001eb385610_HY5	34	-
9	Zm00001eb062460_AP2	34	-
10	Zm00001eb336820_HY5	34	-
11	Zm00001eb050790_AtbHLH15	34	-
12	Zm00001eb070520_AtbHLH15	34	-
13	Zm00001eb265610_AP2	34	-
14	Zm00001eb102450_SEPALLATA3	33	-
15	Zm00001eb068520_AP2	33	-
16	Zm00001eb400120_SEPALLATA3	33	-
17	Zm00001eb311960_KAN	32	-
18	Zm00001eb284010_AG	31	-
19	Zm00001eb076470_VND7	30	(Yamaguchi et al., 2011)
20	Zm00001eb338060_AG	28	-
21	Zm00001eb424040_HY5	28	-
22	Zm00001eb424050_HY5	28	-
23	Zm00001eb145460_AG	26	-
24	Zm00001eb327040_AP1	24	-
25	Zm00001eb213550_AtbHLH9	24	-
26	Zm00001eb120710_AG	23	-
27	Zm00001eb172450_LHY	23	-
28	Zm00001eb023870_At2g44730	23	-
29	Zm00001eb109820_FLP	23	-
30	Zm00001eb406030_AtWRKY53	22	-
31	Zm00001eb203940_AtWRKY22	21	-
32	Zm00001eb199110_At2g44730	21	-
33	Zm00001eb098330_AtWRKY53	21	-
34	Zm00001eb410950_MYB46	21	(Zhong and Ye, 2012)
35	Zm00001eb209480	21	-
36	Zm00001eb118120_AP1	21	-
37	Zm00001eb319350_HSF6	21	-
38	Zm00001eb134890_ZFP7	21	-
39	Zm00001eb310270_AtWRKY53	21	-
40	Zm00001eb312620_Athb-30	21	-
41	Zm00001eb184340	20	-
42	Zm00001eb387370_AtMYB73	20	(Kumar, Campbell, and Turner, 2016)
43	Zm00001eb415760_LHY	19	-
44	Zm00001eb176840_VND7	18	(Yamaguchi et al., 2011)
45	Zm00001eb416800	18	-
46	Zm00001eb229950	18	-
47	Zm00001eb118970_ARF9	18	-
48	Zm00001eb176440_At1g61730	18	-
49	Zm00001eb033570_At2g44730	18	-
50	Zm00001eb260000_VND7	17	(Yamaguchi et al., 2011)

Plain Extremely Randomized Trees Model			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	Zm00001eb355240_AP2	34	-
2	Zm00001eb265610_AP2	34	-
3	Zm00001eb068520_AP2	34	-
4	Zm00001eb062460_AP2	33	-
5	Zm00001eb050790_AtbHLH15	32	-
6	Zm00001eb122740_HY5	31	-
7	Zm00001eb235510_HY5	31	-
8	Zm00001eb336820_HY5	30	-
9	Zm00001eb070520_AtbHLH15	30	-
10	Zm00001eb432100_AP2	30	-
11	Zm00001eb387280_AP2	30	-
12	Zm00001eb327040_AP1	29	-
13	Zm00001eb385610_HY5	29	-
14	Zm00001eb036590_SEPALLATA3	29	-
15	Zm00001eb102450_SEPALLATA3	29	-
16	Zm00001eb317770_SEPALLATA3	29	-
17	Zm00001eb172450_LHY	29	-
18	Zm00001eb284010_AG	28	-
19	Zm00001eb400120_SEPALLATA3	28	-
20	Zm00001eb301590_At5g43700	28	-
21	Zm00001eb109820_FLP	27	-
22	Zm00001eb008690_AGL4	27	-
23	Zm00001eb338060_AG	27	-
24	Zm00001eb311960_KAN	27	-
25	Zm00001eb209480	27	-
26	Zm00001eb319350_HSF6	26	-
27	Zm00001eb023870_At2g44730	26	-
28	Zm00001eb430640_AtERF4	26	-
29	Zm00001eb416800	26	-
30	Zm00001eb310270_AtWRKY53	26	-
31	Zm00001eb118970_ARF9	26	-
32	Zm00001eb406030_AtWRKY53	26	-
33	Zm00001eb427580_GATA-5	25	-
34	Zm00001eb312620_Athb-30	25	-
35	Zm00001eb150840_At5g44210	25	-
36	Zm00001eb229950	25	-
37	Zm00001eb322390_At5g60200	25	-
38	Zm00001eb184340	24	-
39	Zm00001eb374110	24	-
40	Zm00001eb145460_AG	24	-
41	Zm00001eb387370_AtMYB73	24	(Kumar, Campbell, and Turner, 2016)
42	Zm00001eb120710_AG	24	-
43	Zm00001eb134890_ZFP7	24	-
44	Zm00001eb369560_At5g44210	24	-
45	Zm00001eb415770_LHY	24	-
46	Zm00001eb373300_AtbZIP38	24	-
47	Zm00001eb291730_AtERF4	24	-
48	Zm00001eb138380_SHP2	23	-
49	Zm00001eb415760_LHY	23	-
50	Zm00001eb213550_AtbHLH9	23	-

In the maize species, the Hybrid Extremely Randomized Trees Model identified TFs similar to the Hybrid Random Forest Model, including Zm00001eb076470_VND7, Zm00001eb176840_VND7, and Zm00001eb260000_VND7. Additionally, it detected Zm00001eb410950_MYB46, Zm00001eb387370_AtMYB73. However, the plain Extremely Randomized Trees Model only identified Zm00001eb387370_AtMYB73 in the maize species.

Finally, both hybrid and plain AdaBoost Models were applied, with the results displayed in Table 6.17.

Table 6.17: Comparison of the top 50 transcription factors (TFs) predicted to regulate the lignin biosynthesis pathway by the Hybrid AdaBoost Model and the Plain AdaBoost Model on maize Transcriptomic Test Data Set. The frequency of each TF within the top 2000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Hybrid AdaBoost Model			
Rank	Transcription Factor	Freq.	Reference
1	Zm00001eb076470_VND7	34	(Yamaguchi et al., 2011)
2	Zm00001eb176840_VND7	34	(Yamaguchi et al., 2011)
3	Zm00001eb260000_VND7	34	(Yamaguchi et al., 2011)
4	Zm00001eb093920_MYB46	34	(Zhong and Ye, 2012)
5	Zm00001eb355240_AP2	34	-
6	Zm00001eb400120_SEPALLATA3	34	-
7	Zm00001eb070520_AtbHLH15	34	-
8	Zm00001eb311960_KAN	34	-
9	Zm00001eb102450_SEPALLATA3	34	-
10	Zm00001eb036590_SEPALLATA3	34	-
11	Zm00001eb317770_SEPALLATA3	34	-
12	Zm00001eb122740_HY5	34	-
13	Zm00001eb424050_HY5	34	-
14	Zm00001eb424040_HY5	34	-
15	Zm00001eb050790_AtbHLH15	34	-
16	Zm00001eb432100_AP2	34	-
17	Zm00001eb385610_HY5	34	-
18	Zm00001eb068520_AP2	34	-
19	Zm00001eb265610_AP2	34	-
20	Zm00001eb387280_AP2	34	-
21	Zm00001eb336820_HY5	34	-
22	Zm00001eb062460_AP2	34	-
23	Zm00001eb004600_AtbHLH15	34	-
24	Zm00001eb235510_HY5	34	-
25	Zm00001eb213550_AtbHLH9	33	-
26	Zm00001eb284010_AG	33	-
27	Zm00001eb098330_AtWRKY53	33	-
28	Zm00001eb338060_AG	33	-
29	Zm00001eb410950_MYB46	32	(Zhong and Ye, 2012)
30	Zm00001eb327040_AP1	31	-
31	Zm00001eb118120_AP1	30	-
32	Zm00001eb172450_LHY	30	-
33	Zm00001eb359470_AtWRKY53	30	-
34	Zm00001eb145460_AG	29	-
35	Zm00001eb310270_AtWRKY53	28	-
36	Zm00001eb406030_AtWRKY53	28	-
37	Zm00001eb344810_AtWRKY22	28	-
38	Zm00001eb376400_AtWRKY53	27	-
39	Zm00001eb159410_AtWRKY22	27	-
40	Zm00001eb184340	27	-
41	Zm00001eb120710_AG	26	-
42	Zm00001eb358680_AtWRKY22	26	-
43	Zm00001eb041650_AGL15	25	-
44	Zm00001eb344160_AtWRKY53	24	-
45	Zm00001eb203940_AtWRKY22	24	-
46	Zm00001eb134890_ZFP7	22	-
47	Zm00001eb193550_At1g21910	20	-
48	Zm00001eb202570	20	-
49	Zm00001eb312620_Athb-30	19	-
50	Zm00001eb348560	19	-

Plain AdaBoost Model			
Rank	Transcription Factor	Freq.	Reference
1	Zm00001eb355240_AP2	34	-
2	Zm00001eb068520_AP2	34	-
3	Zm00001eb265610_AP2	34	-
4	Zm00001eb062460_AP2	34	-
5	Zm00001eb387280_AP2	34	-
6	Zm00001eb050790_AtbHLH15	34	-
7	Zm00001eb122740_HY5	34	-
8	Zm00001eb231360_AthZIP44	34	-
9	Zm00001eb070520_AtbHLH15	34	-
10	Zm00001eb432100_AP2	33	-
11	Zm00001eb336820_HY5	32	-
12	Zm00001eb235510_HY5	32	-
13	Zm00001eb417610_PIF3	31	-
14	Zm00001eb424050_HY5	31	-
15	Zm00001eb385610_HY5	31	-
16	Zm00001eb102450_SEPALLATA3	29	-
17	Zm00001eb175540_AtbHLH137	29	-
18	Zm00001eb072360	29	-
19	Zm00001eb189510_OBP4	29	-
20	Zm00001eb424040_HY5	29	-
21	Zm00001eb391230_COL4	29	-
22	Zm00001eb338060_AG	29	-
23	Zm00001eb152350	27	-
24	Zm00001eb194220_At1g78080	27	-
25	Zm00001eb023220_COL4	25	-
26	Zm00001eb427650_ATHB-7	24	-
27	Zm00001eb004600_AtbHLH15	24	-
28	Zm00001eb400130_CRC	24	-
29	Zm00001eb327040_AP1	24	-
30	Zm00001eb335690_At2g38090	24	-
31	Zm00001eb284010_AG	24	-
32	Zm00001eb008690_AGL4	24	-
33	Zm00001eb311960_KAN	24	-
34	Zm00001eb317770_SEPALLATA3	24	-
35	Zm00001eb006180	19	-
36	Zm00001eb036590_SEPALLATA3	19	-
37	Zm00001eb400120_SEPALLATA3	18	-
38	Zm00001eb348560	18	-
39	Zm00001eb076470_VND7	18	(Yamaguchi et al., 2011)
40	Zm00001eb209070_AthZIP11	16	-
41	Zm00001eb327140_NF-YB3	16	-
42	Zm00001eb028820_AtMYB59	16	-
43	Zm00001eb138380_SHP2	15	-
44	Zm00001eb051380_At5g39660	15	-
45	Zm00001eb144340_GAI	15	-
46	Zm00001eb066100_At5g66940	15	-
47	Zm00001eb357220_At3g13810	12	-
48	Zm00001eb310270_AtWRKY53	11	-
49	Zm00001eb139600_AtMYB73	11	(Kumar, Campbell, and Turner, 2016)
50	Zm00001eb051900_At4g34610	11	-

Table 6.17 reveals that the Hybrid AdaBoost Model identified TFs similar to the Hybrid Random Forest Model, including Zm00001eb176840_VND7, Zm00001eb260000_VND7, Zm00001eb076470_VND7, Zm00001eb093920_MYB46, and Zm00001eb410950_MYB46, whereas the plain AdaBoost Model detected only two TFs which are Zm00001eb076470_VND7 and Zm00001eb139600_AtMYB73. To compare the performance of the models with a baseline model, Spearman's rank correlation coefficient was applied to the maize transcriptomic test data, similar to the *Arabidopsis* and poplar gene expression analysis. A False Discovery Rate (FDR) correction was applied, and the top 50 TFs were extracted. Table 6.18 displays the top 50 TFs as well as their corresponding ranks.

Table 6.18: Top 50 transcription factors (TFs) that regulate the lignin biosynthesis pathway based on the corrected Spearman correlation coefficient on maize Transcriptomic Test Data Set. The frequency of each TF within the top 2000 predicted regulatory relationships was calculated to represent how many pathway genes it might have inferred. TFs highlighted in red represent true regulators according to current literature, with the corresponding references provided.

Spearman Correlation Coefficient			
Rank	**Transcription Factor**	**Freq.**	**Reference**
1	Zm00001eb112680_AtbHLH62	13	-
2	Zm00001eb055830_FIT1	12	-
3	Zm00001eb426670_AtMYB15	12	-
4	Zm00001eb077700_AtMYB15	12	-
5	Zm00001eb191940_GATA-12	11	-
6	Zm00001eb164490	11	-
7	Zm00001eb318060	11	-
8	Zm00001eb159340_AtWRKY28	10	-
9	Zm00001eb137920_WLIM1	10	-
10	Zm00001eb320470_At3g54390	10	-
11	Zm00001eb015120_AtMYB112	10	-
12	Zm00001eb044660_At5g60200	10	-
13	Zm00001eb154170_AtWRKY69	10	-
14	Zm00001eb350280_ATWRKY33	9	-
15	Zm00001eb388620_AtWRKY40	9	-
16	Zm00001eb294560_AtMYB86	9	-
17	Zm00001eb429870_AtERF1	9	-
18	Zm00001eb393460_At5g04340	9	-
19	Zm00001eb395580_AtMYB112	9	-
20	Zm00001eb223590_At1g74950	9	-
21	Zm00001eb327450_At1g74950	9	-
22	Zm00001eb195420_AtWRKY55	9	-
23	Zm00001eb213800	9	-
24	Zm00001eb286490_ATWRKY33	9	-
25	Zm00001eb273610	9	-
26	Zm00001eb403720_KNAT7	9	(Yu, 2019)
27	Zm00001eb169340_AtbHLH62	9	-
28	Zm00001eb155610	9	-
29	Zm00001eb001720_KNAT7	9	(Yu, 2019)
30	Zm00001eb006180	8	-
31	Zm00001eb153330_WRKY	8	-
32	Zm00001eb074930_AtERF1	8	-
33	Zm00001eb330910	8	-
34	Zm00001eb072200	8	-
35	Zm00001eb157260_SND2	8	(Hussey et al., 2011)
36	Zm00001eb125240_At1g68360	8	-
37	Zm00001eb260850_NST2	8	(Mitsuda et al., 2007)
38	Zm00001eb185160_AtMYB15	8	-
39	Zm00001eb342580_AtMYB55	8	-
40	Zm00001eb269810_NST2	8	(Mitsuda et al., 2007)
41	Zm00001eb417490_AtWRKY42	8	-
42	Zm00001eb335320_At1g68360	8	-
43	Zm00001eb068530	8	-
44	Zm00001eb210520	7	-
45	Zm00001eb326170_At3g49930	7	-
46	Zm00001eb030190_ZAT10	7	-
47	Zm00001eb290350_ATWRKY33	7	-
48	Zm00001eb195770	7	-
49	Zm00001eb154560_AtMYB87	7	-
50	Zm00001eb326170_At3g49930	7	-

The Spearman correlation coefficient identified TFs such as Zm00001eb403720_KNAT7 and Zm00001eb001720_KNAT7, known to be regulators of the lignin pathway (Yu 2019). Additionally, it identified Zm00001eb157260_SND2, Zm00001eb260850_NST2, and Zm00001eb269810_NST2, which are involved in SCW formation (Hussey et al. 2011), (Mitsuda et al. 2007). Although the Spearman correlation coefficient and Hybrid Random Forest Models showed comparable performance, the hybrid models consistently outperformed traditional models in making inferences in the maize species. This highlights the effectiveness of hybrid models in identifying key TFs and pathways in complex biological systems.

Finally, a GRN was built using the best-performing hybrid ML model, the Hybrid Random Forest model, on the maize Transcriptomic Test Data. In the Figure 6.9, the target genes of the LBP are represented by green nodes, while the top 50 TFs with the highest connectivity are depicted by the other nodes based on the frequency count. The light coral nodes specifically denote the true transcription factors involved in the LBP.

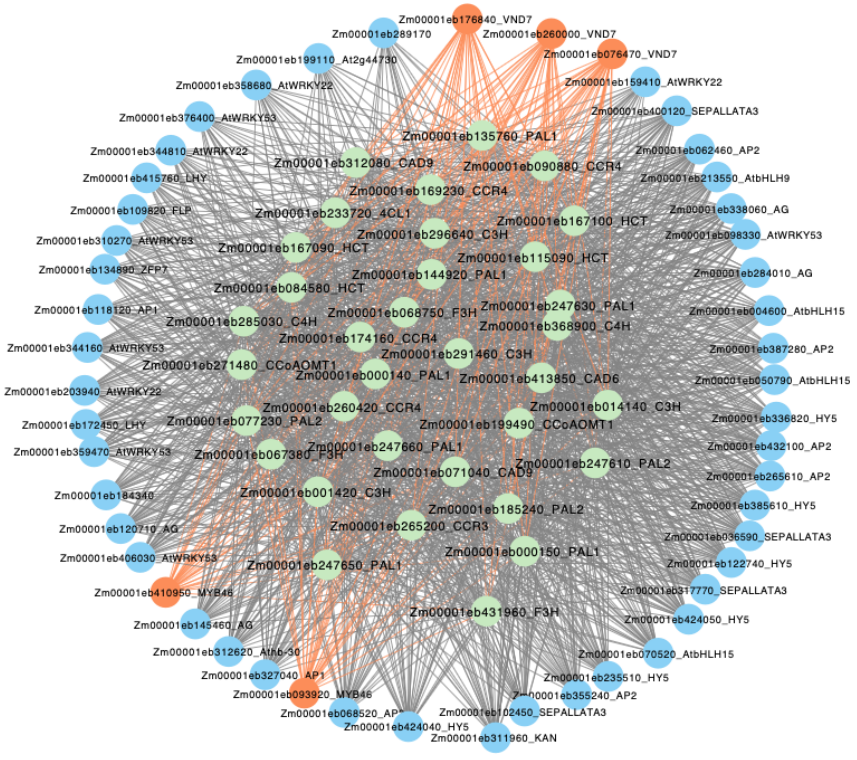

Figure 6.9: Regulatory network generated by the Hybrid Random Forest model on maize Transcriptomic Test Data Set. The green nodes denote the target genes, all the other nodes are the top 50 transcription factors based on the frequency count. The light coral nodes represent the true TFs which involve in the lignin biosynthesis pathway.

7 TF-binding motifs found in the proximal promoter regions of the target genes

Genes are segments of deoxyribonucleic acid (DNA) that can be transcribed to produce message RNA species, which are then translated into proteins to regulate various biological processes, for example, replication, growth and development. The four nucleotides present in DNA are adenine (A), thymine (Wilson et al.), guanine (Hellman and Fried), and cytosine (Robertson et al.), and nucleotide pairings, A=T and C≡G (where each – represents a hydrogen bond), binds the double-stranded DNA (also referred to as sense and antisense DNA strands) together. A TF protein plays an important role in regulating gene expression through binding itself to the proximal promoter region of a target gene and activating or modulating its transcription. Pinpointing the binding site of a TF is a complex but important task in bioinformatics because it can link a TF and its target genes. For instance, this task can be likened to finding a pattern of n-letters, where n usually varies from 6 to 20 nucleotides long, in a large text corpus. The n-letter pattern is the binding site of a TF, and the large text corpus can be the promoter regions of multiple target genes.

The binding site of a TF, which is also referred to as DNA motifs, can be present in both DNA strands and can occur more than once in a strand. In this study, finding the TF-binding site of a TF in candidate target genes would provide additional evidence that the TF may regulate a given target gene. A binding site of a TF is usually represented by a position weight matrix (PWM) as shown in Figure 7.1.

```
>TFmatrixID_0001
A       [0.3651  0.7719   0.8431   0.4837   0.3315   0.6596   0.4972   0.0979   0.0641   0.3922  ]
C       [0.0848  0.0176   0.0353   0.0144   0.0061   0.0029   0.0047   0.0236   0.1464   0.106   ]
G       [0.1309  0.1464   0.0236   0.0047   0.0029   0.0061   0.0144   0.0353   0.0176   0.1365  ]
T       [0.4191  0.0641   0.0979   0.4972   0.6596   0.3315   0.4837   0.8431   0.7719   0.3654  ]
```

Figure 7.1: A position weight matrix (PWM) of the transcription factor (TF) (AT1G63480).Each value in the matrix represents a probability of the nucleotide of adenine (A), thymine (T), cytosine (C) and guanine (G) at a specific position at a DNA motif. This binding site is 10 nucleotides long and the first position has 41.91% probability to a T.

A TF could have more than one motif matrices. A TF may still recognize a motif even if its nucleotides in some positions change, generating variant matrices. If a PWM for a TF is experimentally tested, it is usually referred to as a known PWM, while some PWMs are computationally inferred. These matrices show the probability of a given nucleotide at each position in the matrices. The higher the probability, the greater the likelihood that a particular nucleotide is present at that location in the pattern. These probabilities are used to find TF binding sites in target genes.

7.1 Promoter Region

A promoter is the upstream regulatory sequence of a gene. In the promoter region, the four nucleotides (ATCG) repeat in various pattern. The proximal promoter regions, usually 2 kb nucleotides long upstream of the transcription start sites, of candidate target genes were gathered from the Phytozome Database (Goodstein et al. 2012). Each proximal promoter region was further processed to detect motif binding sites using the MotifLocator program (Thijs et al. 2002). This program uses motif matrix information and sequential data to find the exact locations of a motif of a given TF in the proximal promoter region of each candidate target gene. The program also provides information such as the number of

motifs found in each proximal promoter sequence, and whether the motif is located in the sense or antisense strand.

```
>ATCG00510|ATCG00510.1
ATTCTTCTAGTTATTCCAATAATGTTGATCTTTTAGTTAGCTCCAAGGACATTCGCAATT
TCATATCGGATGACACCTTTTTTGTTAGGGATAGTAATAAGAATAGTTATTCTATATTTT
TTGATAAAAAAAAAAAAATTTTTGAGATTGACAATGATTTTAGTGACCTAGAAAAATTTT
TTTATAGTTATTGTAGTTCTAGTTATCTAAATAATAGATCTAAAGGTGACAACGATCTGC
ACTATGATCCTTACATTAAGGATACTAAATATAATTGTACTAATCACATTAATAGTTGCA
TTGATTCTTATTTTCGTTCTTACATCTGTATTGATAATAACTTTTTAATCGATAGTAATA
ATTTTAATGAAAGTTACATTTATAATTTCATTTGTAGTGAAAGCGGAAAGATTCGTGAAA
GTAAAAATTACAAGATAAGAACTAATAGGAATCGTAGTAATTTAATAAGTTCTAAGGATT
TCGATATAACTCAAAACTACAATCAATTGTGGATTCAATGCGACAATTGTTATGGATTAA
TGTATAAGAAAGTCAAAATGAATGTTTGTGAACAATGTGGACATTATTTGAAAATGAGTA
GTTCAGAAAGAATCGAGCTTTCGATTGATCCGGGTACTTGGAATCCTATGGATGAAGACA
TGGTCTCTGCGGATCCCATTAAATTTCATTCGAAGGAGGAACCTTATAAAAACCGTATTG
ACTCTGCGCAAAAAACTACAGGATTGACTGACGCTGTTCAAACAGGTACAGGTCAACTAA
ACGGTATTCCGGTAGCCCTTGGGGTTATGGATTTTCGGTTTATGGGGGGTAGTATGGGAT
CCGTAGTAGGCGAAAAAATAACTCGTTTGATCGAGTATGCTACCAATCAATGTTTACCTC
TTATTTTAGTGTGTTCTTCCGGAGGAGCACGAATGCAAGAAGGAAGTTTAAGTTTGATGC
AAATGGCTAAAATTTCTTCGGTTTTATGTGATTATCAATCAAGTAAAAAGTTATTCTATA
TATCAATTCTTACATCTCCTACTACCGGTGGAGTGACAGCTAGTTTTGGTATGTTGGGGG
ATATCATTATTGCCGAACCCTATGCCTATATTGCATTTGCGGGTAAAAGAGTAATTGAAC
AAACATTGAAAAAAGCCGTGCCTGAAGGTTCACAAGCGGCTGAATCTTTATTACGTAAGG
GCTTATTGGATGCAATTGTACCACGTAATCTTTTAAAAGGTGTTCTGAGCGAGTTATTTC
AGCTCCATGCTTTTTTTCCTTTGAACACAAATTAAATAAAATAGAACGGTTAGTTTATCA
GAATTAAACGAAAACCCAGAAAAATGCATTTTTCTTTCAAATCATTTTTTTTTATCGATA
TTCTTGTTTACTACTCAGTAAACCTCTATCAACAAGCTAAAAAGTGAATTTTTTTGGGGG
GGAAGTTCAAATTAGACTAGACAAACAAAAAAAAGTTCATTTTCCTCCCTTGCTTGCATA
TGTATAGATAATTCAAATATAGATAGATGCAGATCTATAGAGAGTCTTCCATCTTTTTGC
ATTTCCCGAAAATTCCCTGTTGTTAGATCAGATTCTAATTTCTAATCAATTTTGTAGAAA
TTTGTAATGGAAGAAAAATTTTTCTTTATTAATGACTATTAACTATTATAAGATATTAAG
ACAAAAAGAATAATAAATCTAACAAGGGGATTATGATACATCTAGTTGTAGTTGATTTTG
AAAGATGAATAAGTCCATTTATTTAGTTTGGCTTTTTTTGTACCTATTTTTTTATTCTAT
TTCTATTTCTATTCGGTTCTATTCTATATATTTCTATTAGGTTGTATATTAATATTAGAT
ATATATTTACTTAAAGATACTTAGTATAATTATATAATAGATATAATAGAAATAATAAAA
ATACAAGATATTCTAAGATATCTTTAGAATTCAGAATATAACAATAACAGGTACAAATAT
TAAATTGAGGTACCCATTTT
```

Figure 7.2: An example proximal promoter sequence (2000 nucleotide long) in fasta file format from *Arabidopsis*.The proximal promoter is the upstream regulatory sequence of a gene.

This motif locator program was applied on the *Arabidopsis* Transcriptome Test Data Set 2 which was adopted from Taylor-Teeples's Supplementary Table 2 (Taylor-Teeples et al. 2015). As detailed in Section 3: Multiple OMICS Data Collection, this transcriptomic data set consists of 582 regulatory pairs which are considered to be positive regulatory pairs as they were validated by using Yeast One Hybrid System (Bulyk et al. 1999). Out of the 582 regulatory relationships examined, there are 199 distinct TFs and 44 unique target genes. Position weight matrix information is available for only 141 of these TFs. The top-performing model, the Hybrid Random Forest algorithm, predicted 471 pairs as positive regulatory relationships. The motif locator program further substantiated the predictions

made by the Hybrid Random Forest model, successfully identifying motif locations for 200 out of the 339 predicted pairs which have the position weight matrix information.

8 Transfer Learning

Transfer learning is a ML strategy that leverages knowledge acquired in one domain to enhance performance in a related domain. For plant genomics and bioinformatics, transfer learning allows one to infer gene regulatory relationships in a species when there is no or very limited training data. For example, known gene regulatory relationships are scarce. However, by utilizing known gene regulatory relationships and large quantities of transcriptomic data in one well-studied species like *Arabidopsis*, we can infer gene regulatory relationships in another less-studied species such as a crop or tree species. Transfer learning with CNNs, in particular, can offer significant benefits as these models can automatically learn and extract features from the input data of one species. The learned models can be applied to training data available in another (or a second) species. Note that known gene regulatory relationships and transcriptomic data from the second species are still essential but can be limited. In this case, the learning models from the first species can help fine-tune the models obtained from the second species.

Research has demonstrated the successful application of transfer learning in gene expression data analysis. For example, (Moore et al. 2020) used information from annotated gene expression data of *Arabidopsis* species to classify specialized and general metabolism in tomato plants. Their findings highlight the effectiveness of transfer learning for cross-species analysis.

In this study, a transfer learning approach was implemented to apply knowledge transfer from *Arabidopsis* species to poplar and maize species. This method enabled the models to train on poplar and maize species with fewer training samples than for *Arabidopsis*. The architecture for the transfer learning approach using CNNs is illustrated

in Figure 8.1. Model 1, also known as the base CNN model, was trained on *Arabidopsis* training data. After successful training, the learned parameters for the convolutional layers were transferred to CNN Model 2. CNN Model 2 was trained separately on poplar and maize training data and was tested using their respective testing data.

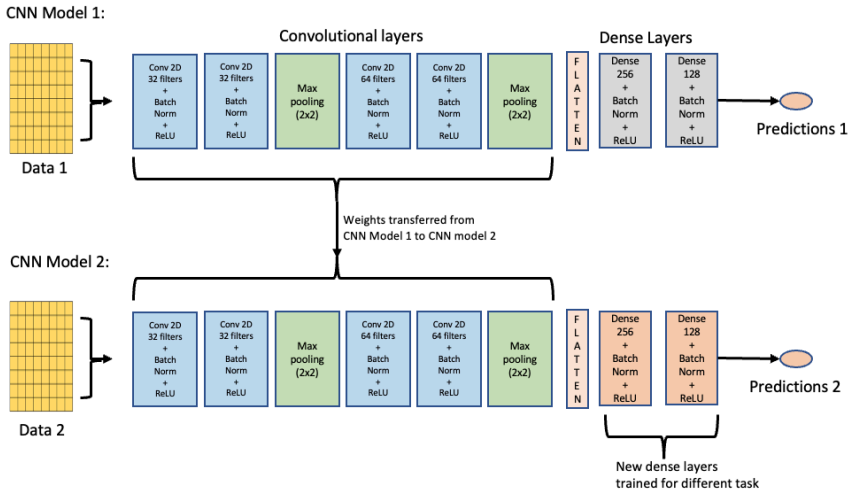

Figure 8.1: Architecture of the transfer learning technique using convolutional neural networks (CNN). Model 1 is trained using data set 1 from a well-studied species, and the learned parameters of the convolutional layers are transferred to model 2. Model 2 has a new dense layer for different tasks and is trained on training data set 2 from a less characterized species. Training data set 2 is usually smaller than Training data set 1.

8.1 Training and Testing data for Transfer Learning:

This section delves into the training and testing data utilized in the transfer learning methodology. The *Arabidopsis* training data, displayed in Table 3.5, includes 2,462 regulatory pairs and 1,253 expression samples per gene; this data set was used to train CNN Model 1. To apply transfer learning in poplar species, 100 positive regulatory pairs were extracted from the poplar training data shown in Table 3.5, which was obtained through homologous mapping from the *Arabidopsis* species. Additionally, 100 negative regulatory pairs were created by randomly pairing the TFs and the target genes which were not present

in positive pairs. To evaluate the transfer learning method, 500 positive regulatory pair samples were similarly extracted from the poplar training data shown in Table 3.5, and a negative set was generated by random pairing of the TFs and target genes which were not present as positive pairs in the poplar training data.

The maize training and testing data for transfer learning was also derived from the maize training data found in Table 3.5. For the maize species, 100 positive regulatory pairs were used, which were obtained through homologous mapping from the *Arab*idopsis species. Likewise, 100 negative regulatory pairs were produced by randomly pairing the TFs and target genes. To assess the transfer learning technique, 500 positive regulatory gene pairs were taken into account, and a negative set was generated in the same way.

8.2 Evaluation of transfer learning

In order to effectively implement transfer learning, it is crucial to train CNN Model 1, also known as the base model. In this study, the base model was trained using *Arabidopsis* training data. CNN Model 1 underwent training on the 80% data of *Arabidopsis* training data for 100 epochs, using Binary Cross-Entropy (BCE) as the loss function, the RMSprop optimizer with a learning rate of 0.00003, and a batch size of 100. Once the model was successfully trained, its performance was evaluated on 20% holdout test data of *Arabidopsis*. CNN Model 1 demonstrated an accuracy of 95.95% and an AUC score of 96.03% for *Arabidopsis* species holdout test data.

The next step was to apply transfer learning to the poplar and maize species using CNN Model 2, which has a similar structure to CNN Model 1, as depicted in Figure 8.1. The base model's weights were transferred to CNN Model 2 and used to train the poplar and maize data separately. The training of CNN Model 2 occurred in two stages: without

fine-tuning and with fine-tuning. In the first stage, the base model weights were transferred directly and frozen to prevent further training; for fine-tuning, the transferred weights were further trained based on the training data specific to CNN Model 2.

Table 8.1 presents the outcomes of CNN Model 2 for poplar and maize species, comparing its effectiveness without transfer learning, with transfer learning, and with fine-tuning. For the poplar species, CNN Model 2 demonstrated an accuracy of 75% and an AUC score of 75.11% without transfer learning. However, the accuracy significantly increased to 81.3% by utilizing transferred weights from CNN Model 1, which was trained on *Arabidopsis* species, and the AUC score rose to 81.35%. Further fine-tuning of these models led to further improvements in performance with an accuracy of 82.4% and an AUC score of 82.45%.

Table 8.1: Performance Metrics of CNN Models for poplar and maize species using regulatory pair test data. The testing for the poplar and maize species consisted of 500 positive pairs which were extracted through homologous mapping, 500 negative pairs which were randomly paired transcription factor and target genes which were not shown as positive pairs.

No.	Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC score
1	Poplar Species CNN Model 2 without Transfer Learning	75	75.24	75.11	79.51	74.98	75.11
2	Poplar Species CNN Model 2 with Transfer Learning	81.3	81.36	81.35	83.61	81.3	81.35
3	Poplar Species CNN Model 2 with Transfer Learning and Fine Tuning	82.4	82.46	82.45	84.63	82.4	82.45
4	Maize Species CNN Model 2 without Transfer Learning	50.5	57.74	50.78	3.38	36.39	50.78
5	Maize Species CNN Model 2 with Transfer Learning	76	76.01	75.99	77.34	75.99	75.99
6	Maize Species CNN Model 2 with Transfer	76.9	77.16	76.87	81.91	76.83	76.87

	Learning and Fine Tuning						

For the maize species, substantial improvement was evident when transfer learning was applied. Without transfer learning, the model's accuracy was only 50.5%, equivalent to random class guessing. In contrast, the model employing transfer learning achieved 76% accuracy and a 75.99% AUC score. Fine-tuning the model further enhanced its performance, with an accuracy of 76.9% and an AUC score of 76.87%. Figure 8.2 presents the ROC curves for the six CNN Models evaluated for poplar and maize species.

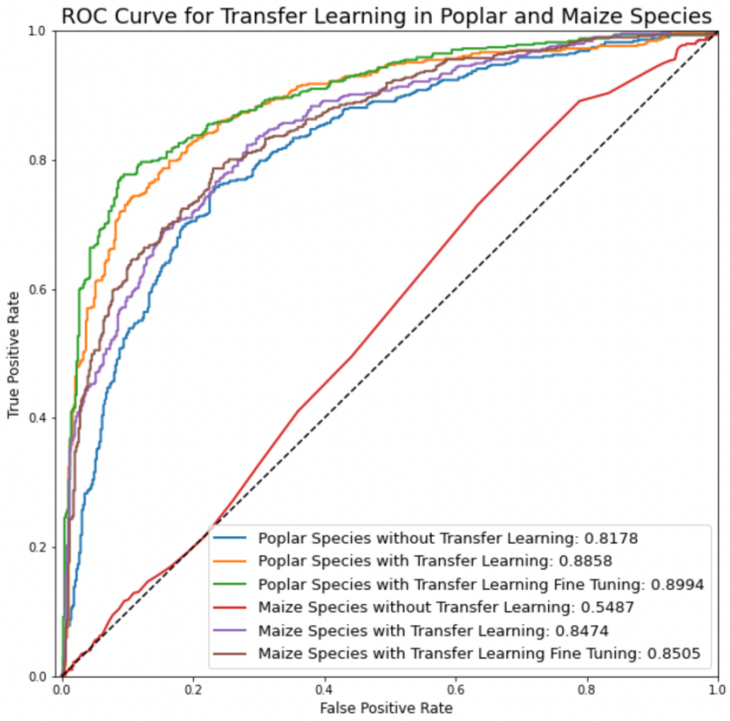

Figure 8.2: Performance comparison of CNN Models with and without transfer learning for poplar and maize species.

The ROC curves demonstrate the improved performance achieved through the incorporation of transfer learning and fine-tuning. For both species, the models employing

transfer learning exhibited a significant enhancement in performance compared to those without transfer learning. Additionally, the fine-tuning process further optimized the models, resulting in higher accuracy and AUC scores.

In this study, the effectiveness of CNN Model 2 on poplar and maize species with and without the application of transfer learning and fine-tuning has been assessed. The results presented in Table 8.1 and Figure 8.2 conclusively demonstrate the substantial benefits of incorporating transfer learning and fine-tuning into the models. For both poplar and maize species, transfer learning dramatically improves model accuracy and AUC scores, while further fine-tuning leads to additional performance enhancements. These findings highlight the importance of leveraging prior knowledge from related tasks or species, as it significantly improves the model's ability to generalize and make accurate predictions. In conclusion, transfer learning and fine-tuning techniques are valuable tools for enhancing the performance of CNN models in the classification of plant species and can potentially be applied to various other domains and tasks to achieve superior results.

9 Discussion

In this study, we addressed the challenges associated with constructing GRNs in plant species by utilizing a combination of ML and ANN approaches. GRNs play a crucial role in understanding the regulation of various metabolic pathways, biological processes, and complex traits in plant growth and response to environmental stimuli. Conventional experimental methods for constructing GRNs are labor-intensive and time-consuming. Therefore, the most efficient approach to construct GRNs is to identify candidate regulatory relationships using *in-silico* methods and then use experiment means to validate these candidate regulatory pairs.

Our research highlights the potential of ML, ANN, and hybrid techniques for accurate GRN prediction using *Arabidopsis* thaliana, poplar, and maize transcriptomic data from the NCBI SRA database. We assessed various ML models and ANN approaches (including FCNs and CNNs), for GRN construction in plant species and evaluated their effectiveness on transcriptomic data. Additionally, we explored hybrid models combining CNN and ML methods to enhance gene regulatory pair predictions. Our findings reveal that hybrid models, such as Hybrid Random Forest, Hybrid Extremely Randomized Trees, and Hybrid Adaboost Models, significantly outperform traditional ML and ANN approaches on holdout test data and real test data, including lignin biosynthesis pathway analysis. Moreover, we developed a program to identify TF-binding motif locations, which effectively detected motifs in *Arabidopsis* species, reinforcing algorithm predictions. However, the unavailability of position weight matrix information for poplar and maize is a limitation of this approach.

One limitation of our study was that only small training data sets with a limited number of gene regulatory pairs are available in *Arabidopsis*, poplar, and maize. The TF-target gene pairs in the training data for less-studied species such as poplar and maize were identified through homologous mapping from *Arabidopsis*. Limited training data presents a significant constraint for supervised learning approaches, where training data is crucial. To address this issue, we utilized transfer learning techniques to explore whether knowledge gained from one species could be transferred to another. We trained a convolutional encoder on *Arabidopsis* species data and then applied the encoder to poplar and maize species data. The results showed a substantial improvement in performance using transfer learning compared to the outcomes obtained without transfer learning. This finding implies that cross-species analysis offers promising potential for future research in GRN construction and prediction.

In conclusion, our study highlights the effectiveness of hybrid models that integrate ML and ANN approaches in predicting GRNs for plant species. These hybrid models exhibit superior performance compared to traditional ML or ANN methods alone. Additionally, the implementation of transfer learning techniques provides valuable insights into cross-species analysis, paving the way for future research and development in the field of GRN prediction and construction.

10 Conclusion

This study effectively demonstrated the potential of combining ML and CNN approaches for constructing GRNs in plant species using transcriptome data from *Arabidopsis thaliana*, poplar, and maize. Hybrid models integrating CNN and ML techniques outperformed traditional ML or ANN methods in predicting gene regulatory relationships. Although CNN methods' potential was not fully realized due to limited training data, transfer learning techniques significantly improved performance, indicating the viability of cross-species analysis for future GRN research.

This study contributes to the growing knowledge of GRN prediction in plant species by showcasing the value of hybrid models and transfer learning. These approaches enhance gene regulatory predictions and offer insights into cross-species analysis, paving the way for future research to improve our understanding of plant growth, responses to environmental stimuli, and regulation of metabolic pathways and biological processes.

11 References

Abadi, Martín, Paul Barham, Jianmin Chen, Z. Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zhang. 2016. 'TensorFlow: A system for large-scale machine learning', *arXiv*, abs/1605.08695.

B. Khojasteh, Hakimeh and Olyaee, Mohammad and Khanteymoori, Alireza. 2021. 'EnGRNT: Inference of gene regulatory networks using ensemble methods and topological feature extraction'.

Bartlett, A., R. C. O'Malley, S. S. C. Huang, M. Galli, J. R. Nery, A. Gallavotti, and J. R. Ecker. 2017. 'Mapping genome-wide transcription-factor binding sites using DAP-seq', *Nature Protocols*, 12: 1659-72.

Benjamini, Y., and Y. Hochberg. 1995. 'Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 57: 289-300.

Boerjan, W., J. Ralph, and M. Baucher. 2003. 'Lignin biosynthesis', *Annu Rev Plant Biol*, 54: 519-46.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30: 2114-20.

Breiman, L. 1996. 'Bagging predictors', *Machine Learning*, 24: 123-40.

Bulyk, M. L., E. Gentalen, D. J. Lockhart, and G. M. Church. 1999. 'Quantifying DNA-protein interactions by double-stranded DNA arrays', *Nature Biotechnology*, 17: 573-77.

Butte, A. J., and I. S. Kohane. 2000. 'Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements', *Pac Symp Biocomput*: 418-29.

Chai, G. H., Y. Z. Kong, M. Zhu, L. Yu, G. Qi, X. F. Tang, Z. G. Wang, Y. P. Cao, C. J. Yu, and G. K. Zhou. 2015. 'Arabidopsis C3H14 and C3H15 have overlapping roles in the regulation of secondary wall thickening and anther development', *Journal of Experimental Botany*, 66: 2595-609.

Chen, Y. F., Y. Li, R. Narayan, A. Subramanian, and X. H. Xie. 2016. 'Gene expression inference with deep learning', *Bioinformatics*, 32: 1832-39.

Choi, S. H., A. T. Labadorf, R. H. Myers, K. L. Lunetta, J. Dupuis, and A. L. DeStefano. 2017. 'Evaluation of logistic regression models and effect of covariates for case-control study in RNA-Seq analysis', *BMC Bioinformatics*, 18: 91.

Cortes, Corinna and Vapnik, Vladimir. 1995. 'Support-vector networks', *Machine Learning*, 20: 273-97.

Danaher, P., P. Wang, and D. M. Witten. 2014. 'The joint graphical lasso for inverse covariance estimation across multiple classes', *J R Stat Soc Series B Stat Methodol*, 76: 373-97.

Deng, W., K. Zhang, V. Busov, and H. Wei. 2017. 'Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways', *PLoS One*, 12: e0171532.

Deng, W., K. Zhang, S. Liu, P. X. Zhao, S. Xu, and H. Wei. 2018. 'JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions', *Bioinformatics*, 34: 3470-78.

Dewey GT, Galas DJ. 2000-2013. *Gene Regulatory Networks.* (Madame Curie Bioscience Database: [Internet]).

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29: 15-21.

Dojer, N., A. Gambin, A. Mizera, B. Wilczynski, and J. Tiuryn. 2006. 'Applying dynamic Bayesian networks to perturbed gene expression data', *Bmc Bioinformatics*, 7.

Eetemadi, A., and I. Tagkopoulos. 2019. 'Genetic Neural Networks: an artificial neural network architecture for capturing gene expression relationships', *Bioinformatics*, 35: 2226-34.

Friedman, J. H. 2001. 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics*, 29: 1189-232.

Geng, P., S. Zhang, J. Liu, C. Zhao, J. Wu, Y. Cao, C. Fu, X. Han, H. He, and Q. Zhao. 2020. 'MYB20, MYB42, MYB43, and MYB85 Regulate Phenylalanine and Lignin Biosynthesis during Secondary Cell Wall Formation', *Plant Physiol*, 182: 1272-83.

Geurts, P., D. Ernst, and L. Wehenkel. 2006. 'Extremely randomized trees', *Machine Learning*, 63: 3-42.

Gillani, Z., M. S. Akash, M. D. Rahaman, and M. Chen. 2014. 'CompareSVM: supervised, Support Vector Machine (SVM) inference of gene regularity networks', *BMC Bioinformatics*, 15: 395.

Goodstein, D. M., S. Q. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D. S. Rokhsar. 2012. 'Phytozome: a comparative platform for green plant genomics', *Nucleic Acids Research*, 40: D1178-D86.

Gunasekara, C., K. Zhang, W. Deng, L. Brown, and H. Wei. 2018. 'TGMI: an efficient algorithm for identifying pathway regulators through evaluation of triple-gene mutual interaction', *Nucleic Acids Res*, 46: e67.

He, Kaiming, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. 'Deep Residual Learning for Image Recognition', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 770-78.

Hellman, L. M., and M. G. Fried. 2007. 'Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions', *Nat Protoc*, 2: 1849-61.

Ho, Tin Kam. 1995. 'Random decision forests', *IEEE*, 1: 278-82.

Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications', *arXiv*, abs/1704.04861.

Hussey, S. G., E. Mizrachi, A. V. Spokevicius, G. Bossinger, D. K. Berger, and A. A. Myburg. 2011. 'SND2, a NAC transcription factor gene, regulates genes involved in secondary cell wall development in Arabidopsis fibres and increases fibre cell area in Eucalyptus', *Bmc Plant Biology*, 11.

Huynh-Thu, V. A., A. Irrthum, L. Wehenkel, and P. Geurts. 2010. 'Inferring regulatory networks from expression data using tree-based methods', *PLoS One*, 5.

Jiao, Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, J. C. Stein, X. Wei, C. S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K. L. Schneider, T. K. Wolfgruber, M. R. May, N. M. Springer, E. Antoniou, W. R. McCombie, G. G. Presting, M. McMullen, J. Ross-Ibarra, R. K. Dawe, A. Hastie, D. R. Rank, and D. Ware. 2017. 'Improved maize reference genome with single-molecule technologies', *Nature*, 546: 524-27.

Jochen Supper, Holger Fröhlich, Christian Spieth, Andreas Dräger, and Andreas Zell. 2007. 'Inferring gene regulatory networks by machine learning methods', *Series on Advances in Bioinformatics and Computational Biology*, Proceedings of the 5th Asia-Pacific Bioinformatics Conference.

Kamiya, T., M. Borghi, P. Wang, J. M. C. Danku, L. Kalmbach, P. S. Hosmani, S. Naseer, T. Fujiwara, N. Geldner, and D. E. Salt. 2015. 'The MYB36 transcription factor orchestrates Casparian strip formation', *Proceedings of the National Academy of Sciences of the United States of America*, 112: 10533-38.

Kelley, Henry J. 1960. 'Gradient Theory of Optimal Flight Paths', *ARS Journal*, 30: 947-54.

Kong, Y. C., and T. W. Yu. 2018. 'A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification', *Scientific Reports*, 8.

Kumar, M., L. Campbell, and S. Turner. 2016. 'Secondary cell walls: biosynthesis and manipulation', *J Exp Bot*, 67: 515-31.

Kumari, S., J. Nie, H. S. Chen, H. Ma, R. Stewart, X. Li, M. Z. Lu, W. M. Taylor, and H. R. Wei. 2012. 'Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery', *PLoS One*, 7.

Luo, W., K. D. Hankenson, and P. J. Woolf. 2008. 'Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information', *BMC Bioinformatics*, 9: 467.

Lyudmyla Yasinska-Damri, Sergii Babichev, Bohdan Durnyak & Tatiana Goncharenko 2022. 'Application of Convolutional Neural Network for Gene Expression Data Classification', *Lecture Notes on Data Engineering and Communications Technologies*, 149.

Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. 2006. 'ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context', *Bmc Bioinformatics*, 7.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models* (New York).

Mitsuda, N., A. Iwase, H. Yamamoto, M. Yoshida, M. Seki, K. Shinozaki, and M. Ohme-Takagi. 2007. 'NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of Arabidopsis', *Plant Cell*, 19: 270-80.

Moore, B. M., P. P. Wang, P. X. Fan, A. Lee, B. Leong, Y. R. Lou, C. A. Schenck, K. Sugimoto, R. Last, M. D. Lehti-Shiu, C. S. Barry, and S. H. Shiu. 2020. 'Within- and cross-species predictions of plant specialized metabolism genes using transfer learning', *In Silico Plants*, 2.

Ohashi-Ito, K., Y. Oda, and H. Fukuda. 2010. 'Arabidopsis VASCULAR-RELATED NAC-DOMAIN6 Directly Regulates the Genes That Govern Programmed Cell Death and Secondary Wall Formation during Xylem Differentiation', *Plant Cell*, 22: 3461-73.

Parry, R. M., W. Jones, T. H. Stokes, J. H. Phan, R. A. Moffitt, H. Fang, L. Shi, A. Oberthuer, M. Fischer, W. Tong, and M. D. Wang. 2010. 'k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction', *Pharmacogenomics J*, 10: 292-309.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12: 2825-30.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. J. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. 2007. 'Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing', *Nature Methods*, 4: 651-57.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Rukhsar, L., W. H. Bangyal, M. S. A. Khan, A. A. A. Ibrahim, K. Nisar, and D. B. Rawat. 2022. 'Analyzing RNA-Seq Gene Expression Data Using Deep Learning Approaches for Cancer Classification', *Applied Sciences-Basel*, 12.

Ruklisa, D., A. Brazma, and J. Viksna. 2005. 'Reconstruction of gene regulatory networks under the finite state linear model', *Genome Inform*, 16: 225-36.

Sergio Peignier, Baptiste Sorin, Federica Calevro. 2021. 'Ensemble learning based gene regulatory network inference', *IEEE*.

Sun, Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian. 2015. 'Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification', *arXiv*.

Tabares-Soto, R., S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodriguez-Sotelo, and C. F. Jimenez-Varon. 2020. 'A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data', *Peerj Computer Science*.

Taylor-Teeples, M., L. Lin, M. de Lucas, G. Turco, T. W. Toal, A. Gaudinier, N. F. Young, G. M. Trabucco, M. T. Veling, R. Lamothe, P. P. Handakumbura, G. Xiong, C. Wang, J. Corwin, A. Tsoukalas, L. Zhang, D. Ware, M. Pauly, D. J. Kliebenstein, K. Dehesh, I. Tagkopoulos, G. Breton, J. L. Pruneda-Paz, S. E. Ahnert, S. A. Kay, S. P. Hazen, and S. M. Brady. 2015. 'An Arabidopsis gene regulatory network for secondary cell wall synthesis', *Nature*, 517: 571-5.

Thijs, G., Y. Moreau, F. De Smet, J. Mathys, M. Lescot, S. Rombauts, P. Rouze, B. De Moor, and K. Marchal. 2002. 'INCLUSive: INtegrated clustering, upstream of sequence retrieval and motif sampling', *Bioinformatics*, 18: 331-32.

Wang, H. Z., U. Avci, J. Nakashima, M. G. Hahn, F. Chen, and R. A. Dixon. 2010. 'Mutation of WRKY transcription factors initiates pith secondary wall formation

and increases stem biomass in dicotyledonous plants', *Proceedings of the National Academy of Sciences of the United States of America*, 107: 22338-43.
Wilson, T. E., T. J. Fahrner, M. Johnston, and J. Milbrandt. 1991. 'Identification of the DNA-Binding Site for Ngfi-B by Genetic Selection in Yeast', *Science*, 252: 1296-300.
Wu, X. D., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. 2008. 'Top 10 algorithms in data mining', *Knowledge and Information Systems*, 14: 1-37.
Yamaguchi, M., N. Mitsuda, M. Ohtani, M. Ohme-Takagi, K. Kato, and T. Demura. 2011. 'VASCULAR-RELATED NAC-DOMAIN 7 directly regulates the expression of a broad range of genes for xylem vessel formation', *Plant Journal*, 66: 579-90.
Yang, C. Y., Z. Y. Xu, J. Song, K. Conner, G. V. Barrena, and Z. A. Wilson. 2007. 'Arabidopsis MYB26/MALE STERILE35 regulates secondary thickening in the endothecium and is essential for anther dehiscence', *Plant Cell*, 19: 534-48.
Yilmaz, A., M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold. 2011. 'AGRIS: the Arabidopsis Gene Regulatory Information Server, an update', *Nucleic Acids Res*, 39: D1118-22.
Yu, Y. Q. 2019. 'OsKNAT7 Bridges Secondary Cell Wall Formation and Cell Growth Regulation', *Plant Physiology*, 181: 385-86.
Zhao, C. S., U. Avci, E. H. Grant, C. H. Haigler, and E. P. Beers. 2008. 'XND1, a member of the NAC domain family in Arabidopsis thaliana, negatively regulates lignocellulose synthesis and programmed cell death in xylem', *Plant Journal*, 53: 425-36.
Zhong, R. Q., and Z. H. Ye. 2012. 'MYB46 and MYB83 Bind to the SMRE Sites and Directly Activate a Suite of Transcription Factors and Secondary Wall Biosynthetic Genes', *Plant and Cell Physiology*, 53: 368-80.
Zhou, J. L., R. Q. Zhong, and Z. H. Ye. 2014. 'Arabidopsis NAC Domain Proteins, VND1 to VND5, Are Transcriptional Regulators of Secondary Wall Biosynthesis in Vessels', *PLoS One*, 9.