Awards issued prior to January 20, 2025, were funded under previous Administrations and may not reflect the priorities and policies of the current Administration.

United States Department of Agriculture
Project Initiation

Project Data for Accession Number 1024681

Project Title

Advancing Machine Learning and Industry 4.0 for the Forest Products Industries based on Automated Data Fusion and Total Quality data Management (TQdM)

Project Details

Sponsoring Institution National Institute of Food and Agriculture	**Project Status** COMPLETE
Funding Source NIFA Formula	**Grant Year** 2021
Grant No.	**Cumulative Award Amt.** N/A
Proposal No.	**Multistate No.** (N/A)
Project Start Date Oct 27, 2020	**Project End Date** Sep 30, 2025
Program Code	

Project Director

Young, T.

Recipient Organization

SAES - UNIVERSITY OF TENNESSEE

1331 CIRCLE PARK DR

KNOXVILLE, TENNESSEE 37916-3801

Performing Department

Forestry, Wildlife & Fisheries

Non Technical Summary

Scientific knowledge will be improved by addressing the problem of information incompleteness and automated algorithms for data quality improvement in the context of Industry 4.0. Research in data science for the sustainable biomaterials forest products industries is not well documented. The development of software for 'real-time automated data fusion,' 'advanced algorithms for data quality improvement' that include predictive 'machine learning algorithms' will not only benefit the scientific community, but directly benefit industries manufacturing forest products and sustainable biomaterials. This benefits public welfare by improving the efficient use of fiber from precious forest resources and sustaining labor markets.This project addresses three of the five legislatively mandated forestry research areas with 'science of integration,' improved systems for assessing uncertainty, and improved decision-making,' with technological advancements to improve productivity in the sustainable biomaterials forest products industries. The authors are uniquely qualified to successfully implement the proposed research given prior research

United States Department of Agriculture
Project Initiation

(André et al. 2008, Clapp et al. 2008, Kim et al. 2012, André and Young 2013, Young et al. 2014, Carty et al. 2015, Young et al. 2015, Zeng et al. 2016, Tian et al. 2018, Young et al. 2019). The proposed research will narrow the large gap that exists between laboratory-based research in this field and successful application by industry, i.e., recall Liao et al. (2017) has indicated this gap as large 90%.This research also addresses the problem of information incompleteness and automated algorithms for TQdM. However, research on 'data quality science' for the sustainable biomaterials and forest products industries is not well documented. Given this need, we propose to develop software systems for 'real-time automated data fusion,' 'advanced algorithms for data quality improvement,' for use in machine learning algorithms. The proposed research is directly aligned with contemporary scientific views on the use of AI and big data for prediction of manufacturing system outputs.

Animal Health Component

Animal Health Component: 0%

Research Effort Categories

Basic	40%
Applied	40%
Developmental	20%

Classification

Knowledge Area (KA)	Subject of Investigation (SOI)	Field of Science (FOS)	Percent
123	0650	2080	100%

Knowledge Area

[123] Management and Sustainability of Forest Resources

Subject Of Investigation

[0650] Wood and wood products

Field Of Science

[2080] Mathematics and computer sciences

Keywords

Machine Learning, AI, Data Fusion, Predictive Models, Total Quality data Management (TQdM)

Goals / Objectives

This project addresses three of the five legislatively mandated forestry research areas with 'science of integration, improved systems for assessing uncertainty, and improved decision-making,' with technological advancements to improve productivity in the sustainable biomaterials forest products industries. The overall goal of the project is to to develop automated machine learning algorithms that will fuse multiple digital databases in manufacturing settings. After data fusion, algorithms will assess and improve (if necessary) the data quality of the fused databases. Predictive machine learning algorithms will be used to predict real-time final product properties, and diagnosis sources of

variation in manufacturing processes that lead to waste of wood fiber, inefficient energy use, uneconomical use of chemical additives, and overall higher manufacturing costs.The following objectives will be fulfilled to satisfy the research project:1. Develop automated algorithms with software interfaces that will allow for digital data integration and data fusion of multiple databases;2. Develop advanced algorithms for data quality assessment and improve Total Quality data Management (TQdM) protocols;3. Develop an advanced machine learning system using at least seven algorithms that include Bayesian methods with software interfaces;4. Validate the entire system by working with the sustainable biomaterials and forest products industries.

Project Methods

METHODSAutomated digital data integration and data fusion. -- An integrated database that fuses the process data with the lab destructive test data will be developed. Most databases in the forest product industries are commercial where the data are not encrypted and are accessible without specialized decryption code (e.g., Microsoft SQL, OSIsoft PI, Sytec IPS B.V.). After integration, the databases will be stored as comma separated variable (.csv) files which are easily convertible to almost any commercial software platform.Algorithms for data quality assessment. -- In the study by Zeng et al. (2016), maximum-likelihood based imputation using the expectation-maximization (EM) algorithm and multiple imputation (MI) using Markov Chain Monte Carlo (MCMC) simulation were used to improve data quality which achieved improved predictive analytics. MI imputation methods replace each missing value with a set of plausible values that represent the uncertainty of the true value. A direct approach for MI is MCMC (Schafer 1997). In general, MCMC is a collection of techniques for obtaining pseudorandom draws from a probability distribution. In MCMC simulation, one constructs a Markov Chain long enough to establish a stationary distribution, which is the distribution of interest. Zeng et al. (2016) found that better predictive analytics were achieved from imputed datasets than from non-imputed datasets.Develop a machine learning system using at least seven algorithms. -- Mixed stepwise regression and genetic algorithms (GA) will be used for variable preselection in model development. Genetic Algorithms (GA) methods are used in AI and applications extend to engineering design, medical, robotics, and manufacturing applications (Koza 1993, Ansari and Hon 1997, West and Honkanen 2004, Hicks 2003, Ho and Ping 2003, Ozturk and Ozturk 2003, and others). A survey paper by Thengade and Dondal (2012) explains the GA algorithm in general terms, i.e., "the input to the GA is a set of potential solutions encoded as a metric called a 'fitness function' where the GA evaluates each candidate according to the fitness function activity (the candidate with good fitness has high chances to get selected)." The main steps of the GA are selection, crossover, and mutation, which are iterative until a set of variables are selected relative to the fitness function.After variable preselection, predictive models will be built using seven techniques after variable preselection. The seven algorithms will be neural networks (NN), boosted regression trees (BRT), bootstrap forest (BSF), Bayesian additive regression trees (BART), partial least squares (PLS), decision trees (DT), and multiple linear regression (MLR).A voting system will be used for 10-fold cross validation outcomes to determine the best predictive algorithm based on the normalized root mean squared error for prediction (NRMSEP). The 80% training and 20% testing (or validation) rule will be applied for validation. The system will be field-tested and validated at three sustainable biomaterials or forest products manufacturing plants.Given that discussion of common regression methods such as multiple linear regression (MLR), partial least squares (PLS), and neural networks (NN) would be extensive, these methods are well documented and are not defined in this proposal for the sake of brevity (Sewell and Chen 2015, and many other sources). 'Tree-based' regression methods have had good performance in the author's previous studies. Regression tree methods partition the data space into homogeneous sub-regions using a variety of techniques. Regression tree models have the advantage of high explanatory value and stability. Regression trees (RT) identify a hierarchy of interactions, and some unknown interactions, that allow for an improved understanding of the interrelationships effecting the dependent variable. 'Tree Boosting' relies on the philosophy that a small number of simple trees of weak learners that are combined as one model outperforms the predictions of one large RT (Friedman 2001, Schapire 2003, Elith et al. 2008). 'Boosting' builds small one to three node trees sequentially as an ensemble to improve predictions. The result grows new trees by accommodating observations that the existing ensemble predicts poorly (Elith et al. 2008). 'Bootstrap forests' (BSF) provides a useful method for estimating other percentiles of the data. The BSF approach was helpful in estimating the interval of lower percentiles of wood composites in a study by Edwards et al. (2011).Bayesian

additive regression trees (BART) use the same approach as BRT in that trees in the ensemble are grown sequentially to reduce the generalized error of prediction. As Chipman et al. (2010) noted, "the entire model is regularized so that no one tree dominates the prediction of the response surface." Unlike BRTs, tree growing assumes that the parameters of the probability density function of Y are estimated under Bayesian framework (Chipman et al. 2010). BART is a sum-of-trees ensemble which has a greater ability to capture interactions and non-linearity and additive effects among variables (Kapelner and Bleich 2017).