A Method for Sensor Reduction in a Supervised Machine Learning Classification System

Niko Murrell, Ryan Bradley, Nikhil Bajaj, *Member, IEEE,* Julie Whitney, and George T.-C. Chiu, *Senior Member, IEEE*

***Abstract*—Smart devices employing interconnected sensors for feedback and control are being rapidly adopted. Many useful applications for these devices are in markets that demand cost-conscious solutions. Traditional machine learning based control systems often rely on multiple measurements from many sensors to achieve performance targets. An alternative method is presented that leverages a time series output produced by a single sensor. By using domain expert knowledge, the time series output is discretized into finite intervals that correspond to the physical events occurring in the system. Statistical measures are taken across these intervals to serve as the features to the machine learning system. Additional features that decouple key physical metrics are identified, improving the performance of the system. This novel approach requires a more modest data set, and does not compromise performance. The resulting development effort is significantly more cost-effective than traditional sensor-classification systems, not only due to the reduced sensor count, but also due to a significantly simplified and more robust algorithm development and testing step. Results are presented with the case study of a media-type classification system within a printing system, which was deployed to the field as a commercial product.**

***Index Terms*—expert-based systems, integrated design, system-level design, embedded software, sensor systems and applications, machine learning**

I. Introduction

Sensors are rapidly decreasing in cost while performance and accuracy increase. Consequently, many electromechanical devices have incorporated sensor-enabled control schemes. Recently, machine learning algorithms have begun to leverage this trend to enable new functionality. Sensed information may be used to generate input features for algorithms that enable proactive diagnostics, system-awareness, and other more complex tasks such as classification. Concerns arise when the number of sensors and the capability of individual nodes are constrained due to cost or other associated factors like computation time and memory footprint. Previous efforts to address this concern have focused on a reduction of computational requirements during both the training and classification phases of embedded supervised machine learning algorithm development [1]. Methods attempting to minimize the number of features required for classification also exist; these may be used to reduce the number of sensors necessary for a given task.

N. Murrell is with Ethicon, Inc., previously with Lexmark.
R. Bradley was previously with Lexmark.
N. Bajaj and George T.-C. Chiu are with Purdue University.
J. Whitney is with the University of Kentucky, previously with Lexmark.

This work presents a novel method to reduce the number of sensors required for a supervised machine learning classification system. Expert knowledge of expected sensor output variation as a function of intrinsic properties, extrinsic properties, and uncontrollable external factors is used to establish a unique feature set that sufficiently decouples otherwise inseparable classes. The system design and control system were concurrently tuned to elicit distinct dynamic responses within predefined temporal regions of a continuous data stream. The analog data was discretized into several distinct zones of interest corresponding to the sensors response to different dynamical processes. A unique difference method allowed the learning algorithm to extract additional useful information from the confounded data set. This methodology is validated by a case study of a print media classifier system developed for a commercial laser printer, which was manufactured and deployed at a large volume. The resultant classification success exceeded that of embodiments using multiple sensors with only a single sensor. Finally, the implications of this design methodology and advantages over a traditional data-driven classification system are discussed.

II. Background

The goal of simplification of multi-sensor systems by harvesting more independent features from a reduced sensor set relies on modification of the measured object usually based on time or geometry. There are numerous studied methods for dimensionality reduction and representation of time series data. General dimension-reduction and re-representation methods include model-based techniques such as those using hidden Markov models [2], [3]. A second class of methods have attempted to reformulate the data with interpolative or regression methods such as piecewise linear (PLA) [4] or piecewise polynomial (PPA) [5] approximations. Another group of methods uses a symbolic representation optimized with certain constraints such as symbolic aggregated approximation (SAX). Still other methods use transforms such as discrete Fourier [6] or discrete cosine transforms or wavelet systems [7], [8]. Although these methods are largely designed for use on general, potentially multi-dimensional time series, they are frequently tested, presented, and verified on application-specific data from medical data [8] to faults in mechanical gear systems [9].

Once the transformation has been performed, classification training and evaluation can occur. Possible algorithms include 1-nearest neighbors (1NN) or k-nearest neighbors (kNN) [10],

which demonstrated considerable success when implemented with representations like SAX in combination with dynamic time warping [11]. More sophisticated methods such as neural networks, multi-layer perceptrons [12], Bayesian networks [13], support vector machines [14] and decision trees [15] have also been used with success and represent alternative design options. Some methods use information from a transformation, such as warping distance, as an additional feature and integrate this into the classification method [16]. In each case, the features used to train these systems are selected to be as orthogonal as possible and the quality of the resulting algorithm is, amongst other things, a function of that orthogonality.

Often, the system can not be easily simplified, and hardware with embedded supervised machine learning systems is designed using a complex network of various sensors. In theory, this extra data enables the designer to build and test a robust algorithm since a network of sensors can be selected to maximize feature orthogonality. This can lead to a temptation to deploy more sensors and computational resources than is strictly necessary. In industries where customers are highly sensitive to product cost, such as office printing, the strategy is often to deploy a single sensor to partly meet design needs. These attempts have included using a set of electrodes to take electrical measurements of media [17], [18], a camera to measure surface roughness [19] or an ultrasonic sensor to determine media density [20]. The desired combination would result in an inflated infrastructure that is limited in practice by cost and effort. The result is products that fall short of ideal performance in order to maintain competitive cost structures. Although a reduction of algorithm memory and computational requirements is possible [1], modifying the algorithm does not necessarily reduce the cost or footprint of sensing (as opposed to computational) hardware.

III. Methodology

In concurrently developed physical systems, the designer has access to significantly more information about the situation than is often available with analyzing time series data in a general case. Time series data output by a single sensor may contain information about multiple physical quantities due to system dynamic behavior. Therefore, multiple physical quantities do not always need to be measured by the same number of physical sensors. The designer has an opportunity to tune the hardware to produce a time series output from a single sensor and then discretize the output with domain expert knowledge to produce multiple features while preserving orthogonality. This results in a system with fewer sensor nodes and a lower associated cost.

Consider the case of a least-squares support vector machine (LS-SVM) [21], [22] deployed in an embedded classification, solving a multiclass problem (e.g. determine if a presented set of features belongs to which one of several distinct sets). The goal is to take as input a vector $x \in \mathbb{R}^{n_f}$, where n_f is the number of features used for classification, and produce an output $y(x)$ which represents the classifier output. Given $x_k \in \mathbb{R}^{n_f}, k = 1, 2, \ldots, N$ are the feature vectors corresponding to N training examples and y_k are the corresponding true classes (in this case $y_k = +1$ if the measurement belongs to a set and $y_l = -1$ if it does not), the classification algorithm is trained by solving the following optimization problem to determine a best separating hypersurface defined through a nonlinear mapping. Some interpretations of the LS-SVM and other SVMs make assumptions about the variables being independent and identically distributed random variables. While we cannot make this claim for this dataset due to temporal correlation, SVM-type algorithms can still work well in practice as long as the combination of features can provide sufficient separation. In the implementation section, we discuss the distribution of the selected input features, and it can be observationally inferred that an SVM might work well given a geometric rather than probabilistic interpretation of SVM methods.

$$\text{minimize } J_P(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 \tag{1}$$

$$\text{subject to } y_k[w^T \varphi(x_k)_b] = 1 - e_k, k = 1, \ldots, N \tag{2}$$

where the classifier takes the form: $y(x) = \text{sgn}[w^T \varphi(x) + b]$, and $\varphi(x_k)$ is a mapping to a (often) higher dimensional space. In practice the classifier is usually solved for in the dual space, the space of Lagrange multipliers of the constraints, α_k (for $k = 1, 2, \ldots, N$). b is a scalar bias offset term. γ is a regularization parameter that can be used to control overfitting vs. under-fitting behavior, but was set as 1. $w \in \mathbb{R}^{n_f}$ is a vector of weights that, along with the mapping $\varphi(x_k)$ helps to define the decision hypersurface. The dual space classifier takes the form:

$$y(x) = \text{sgn}[\sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b] \tag{3}$$

$K(x, x_k) = \varphi^T(x)\varphi(x_k)$ is a Kernel function (a nonlinear mapping that allows additional flexibility in the classification function). Both the dual space classifier and the solution of the classifier optimization problem can be addressed by considering the Karush-Kuhn-Tucker (KKT) conditions for optimality:

$$\mathbf{w} = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k),$$

$$\sum_{k=1}^{N} \alpha_k y_k = 0,$$

$$\alpha_k = \gamma e_k, \forall k = 1, 2, ..., N,$$

$$y_k[\mathbf{w}^T \varphi(x_k) + b] - 1 + e_k = 0, \forall k = 1, 2, ..., N.$$

This allows assembly of the following matrix equation to solve the KKT system:

$$\begin{pmatrix} 0 & \mathbf{y}^T \\ \mathbf{y} & \mathbf{\Omega} + \frac{\mathbf{I}}{\gamma} \end{pmatrix} \begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{1}_v \end{pmatrix} \tag{4}$$

where $\Omega_{kl} = y_l y_l \varphi(x_k)^T \varphi(x_k) = y_k y_l K(x_k, x_l)$, with $k, l = 1, \ldots, N$. At this point the (nonsparse) matrix equation can be solved for α and b using standard methods (LU factorization, etc.). The Kernel function can take a number of different forms, of which $K(x, x_k) = x_k^T x$ (linear), $K(x, x_k) =$

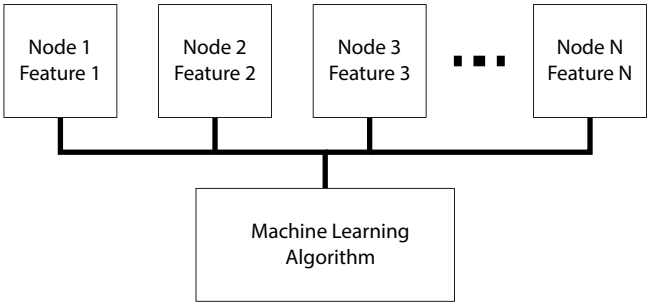


Fig. 1. Traditional method for enabling feature-based decision making capability on an existing device. The final classification algorithm is a function of N features, represented by N nodes.

$(x_k^T x + \tau)^d$ (polynomial), and $K(x, x_k) = \exp\left(-\frac{||x-x_k||_2^2}{\sigma^2}\right)$ (radial basis function) are common examples. In this work, only polynomial classifiers are considered. This is due to the application requirements of processing power and program memory space, constrained to use the algorithm of [1].

Typically, $y(x) = +1$ would yield a prediction that x belongs in one set, and $y(x) = -1$ would correspond to the other complementary set. However, in some cases, including media classification in a printer, there are areas of the feature space that for some comparisons make no difference (there are cases in which mistakes in classification cause less of a problem for downstream processes). Specifically, one can have some errors in classification that are acceptable to downstream processes, and some that should be weighted more heavily. This idea was discussed and formulated into the training of a multiclass SVM problem and described in detail in [23]. The solution method for the system is the same. In this work, the result associated with each classification is accordingly either an incorrect classification, an incorrect (but acceptable) error, or a correct classification. An acceptable error is simply one that is tolerable to the downstream processes.

In order to create a multiclass classification system, the different classes are separated into complementary groups and evaluated in a one vs. all sense [22] (other options exist, but one vs. all is the encoding used in this work); if there are three classes, then there are three classifiers, each of which evaluates whether the data belongs in one set, or alternatively, all of the other sets. As mentioned before, selection of the features that comprise the feature vector are critical to classifier performance. The focus of this work is the design of the features and corresponding sensors and mechanical elements needed in order to achieve good performance while minimizing training data and overall cost.

A traditional approach, shown in Figure 1 places the burden of the system on the sensor nodes themselves. In this example, a feature contributing to the classifier has a one-to-one relationship to the number of required sensor nodes. The proposed approach illustrated by Figure 2 puts the burden of the system on the domain expert knowledge and the temporal output of a single node. The domain expertise is used to partition the measurement time series $m(t)$ (in implementation, this is most likely a sampled time series) into discrete intervals, such that

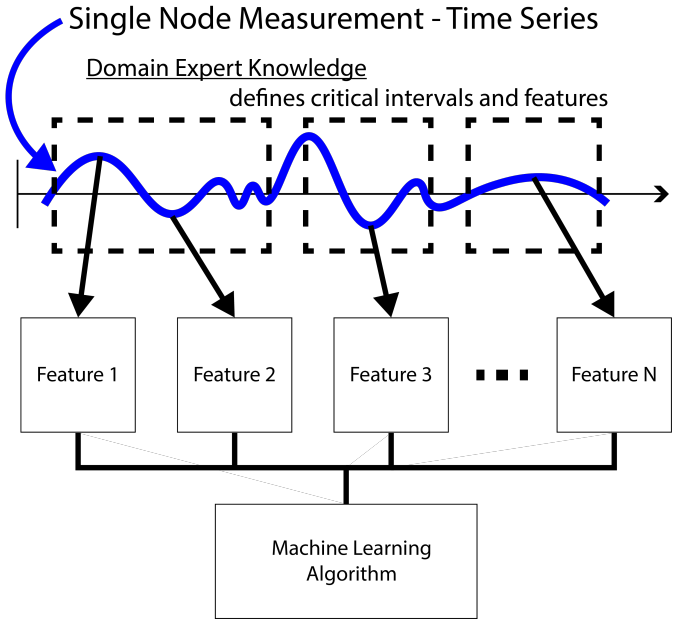


Fig. 2. The proposed approach uses the system knowledge of co-designed hardware to pull multiple features out of a single time series of data.

$$m(t) = \begin{bmatrix} x(t_1,t_2): & [\Psi_{t_1,t_2}], \\ x(t_2,t_3): & [\Psi_{t_2,t_3}], \\ \vdots & \\ x(t_{N-1},t_N): & [\Psi_{t_{N-1},t_N}] \end{bmatrix}$$

Here, the time intervals $[(t_1, t_2), (t_2, t_3), \ldots, (t_{N-1}, t_N)]$ correspond to known physical events in the system and $[x(t_1, t_2), x(t_2, t_3), \ldots, x(t_{N-1}, t_N)]$ is the set of discrete measurement intervals. Ψ is a set of statistical measures (mean, variance, skewness, range, minimum, maximum, etc.) taken within the corresponding measurement interval to describe the interval under inspection.

The classifier is trained on data that is of the form (y_k,x_k). Ideally, $x_k = \phi_k$, where ϕ_k is the set of intrinsic physical properties in the system ($\phi_k = [\phi_1, \phi_2, \ldots, \phi_{N_{pi}}]_k^T \in \mathbb{R}^{N_{pi}}$). N_{pi} represents an ideal set of orthogonal intrinsic properties. $\Psi \subseteq \phi_k$. Simply put, ideally, the sets to be classified are well separated by a measurement of some direct, relevant intrinsic physical property and have good orthogonality. In the practical case, this is not so. Every measurement is a function of both the intrinsic property being measured and the properties of the physical system involved in that measurement. These properties include the structure of the system and its operation, which are controllable by the system designer, and known environmental factors which may not be controllable by the designer. Considering the form of the constructed intervals and corresponding statistical measures, the training data examples

x_k are such that

$$x_k = [f_1(\phi_k, Y_1, Z_k), \\ f_2(\phi_k, Y_2, Z_k), \\ \vdots \\ f_N(\phi_k, Y_N, Z_k)]$$

Here, $(f_1, f_2, \ldots, f_N)$ are nonlinear functions of the arguments: ϕ_k, the intrinsic physical properties; $Z_k \in R^{N_{pe}}$ which are known, quantifiable extrinsic system properties that influence the measurement (N_{pe} is the number of extrinsic properties affecting measurements); and $(Y_1, Y_2, \ldots, Y_N)$, which are uncontrollable external factors that are a function of the hardware design.

In the case of systems where measurements taken in different intervals are coupled, taking the difference between two functions can help to train the classifier with independent information about system interactions and decouple external factors that influence the measurement. This can be justified with a brief expansion analysis. Given two functions f_i and f_j, the Taylor series expansions can be taken about a nominal operating point as

$$f_i(\phi_k, Y_i, Z_k) = \frac{\partial f_i}{\partial \phi_k}\Delta\phi_k + \frac{\partial f_i}{\partial Y_i}\Delta Y_i + \frac{\partial f_i}{\partial z_k}\Delta Z_k + C_i \quad (5)$$

$$f_j(\phi_k, Y_j, Z_k) = \frac{\partial f_j}{\partial \phi_k}\Delta\phi_k + \frac{\partial f_j}{\partial Y_j}\Delta Y_j + \frac{\partial f_j}{\partial z_k}\Delta Z_k + C_j \quad (6)$$

Taking the difference yields

$$f_i(\phi_k, Y_i, Z_k) - f_j(\phi_k, Y_j, Z_k) = \\ \left(\frac{\partial f_i}{\partial \phi_k}\Delta\phi_k + \frac{\partial f_i}{\partial Y_i}\Delta Y_i + \frac{\partial f_i}{\partial Z_k}\Delta Z_k + C_i\right) - \\ \left(\frac{\partial f_j}{\partial \phi_k}\Delta\phi_k + \frac{\partial f_j}{\partial Y_j}\Delta Y_j + \frac{\partial f_j}{\partial Z_k}\Delta Z_k - C_j\right) = \\ \underbrace{\Delta\phi_k\left(\frac{\partial f_i}{\partial \phi_k} - \frac{\partial f_j}{\partial \phi_k}\right)}_{\Delta\phi_k = 0 \text{ for same } k} + \frac{\partial f_i}{\partial Y_i}\Delta Y_i - \frac{\partial f_j}{\partial Y_j}\Delta Y_j + \\ \underbrace{\Delta Z_k\left(\frac{\partial f_i}{\partial Z_k} - \frac{\partial f_j}{\partial Z_k}\right)}_{\Delta Z_k = 0 \text{ for same } k} + \underbrace{C_i - C_j}_{\text{constant}}$$

For the same training example, $\Delta\phi_k = 0$. The same is true for ΔZ_k. Therefore, the only remaining terms are those that include ΔY_i and ΔY_j, the associated partial derivatives, and the difference of the offset constants. This new feature, $f_i - f_j$, is solely a function of ΔY_i and ΔY_j, which are functions of certain fixed extrinsic system properties. This information can be learned by the classifier and improve classification performance.

IV. Case Study and Implementation

This case study applies the proposed approach to a commercial color laser (electrophotographic) printer intended for shared office use in a managed print services environment. Most laser printer users do not check or adjust the media type settings. Additionally, only a fraction of users that do adjust the media settings do so correctly. Incorrect settings on these devices may cause problems for both the customer and the manufacturer. To address this issue, an inexpensive sensor system and embedded machine learning algorithm were implemented to classify media without user input. The printer control system adjusted device parameters based on this media classification.

A single inexpensive optical sensor consisting of a paired LED and phototransistor was mounted within the printer media path. The sensor output a continuous data stream corresponding to the amount of light transmitted by in-process media. A simple model of the sensor was developed and, based upon this, system hardware and controls were tuned to generate an information-rich data stream by leveraging the dynamic response of media to control system inputs. The printer generated features from this data stream for each sheet of media. A broad population of standard office media with varied intrinsic properties, φ_k, existing along a continuum were sorted into 1 of 5 distinct classes: light, normal, heavy, card stock, and transparency. This dataset was used to generate an embedded machine learning algorithm that used these features to determine media class in near real time. Printer process parameters and system controls were adjusted based upon this prediction. The final embodiment significantly reduced overall cost, complexity, and system footprint when compared to traditional implementations and is described in greater detail in [24].

A cross section of the printer media path is shown in Figure 3. The highlighted region contains a section view of the sensor and the surrounding printer hardware including upstream feed rollers, media guides, and downstream feed rollers. The electrical design schematic for the optical sensor system is shown in Figure 4. Nominal circuit values were tuned to adjust the sensor gain, response, and sensitivity. The resulting full scale range of the data set was maximized for the population of expected media and maximum separation between media classes was achieved. Calibration was performed to compensate for system gain and offset errors.

The sensor outputs a continuous data stream corresponding to the amount of infrared light transmitted by the in-process media. This output is a highly coupled function of many confounded factors including intrinsic media properties (e.g., media basis weight, media roughness, media thickness, etc.) φ_k, extrinsic system properties (e.g., LED intensity, media speed, media input source, phototransistor sensitivity, feed roller velocities, media shape and offset, LED and phototransistor directionality, etc.) Z_k, and uncontrollable external factors (e.g., relative humidity, temperature, etc.) Y_i. Figure 5 depicts how variability caused by these confounded factors impacts the measurement for a single media. 20 measurements for a normal weight office media are shown. The signal varies substantially from sheet to sheet and within a given sheet. Sensor output for a given media may vary as much as 20% of the sensors full scale range at a given process point. This is primarily a function of intrinsic media properties, φ_k. Within a given sheet, the sensor output may vary as much as 60% of the sensors full scale range. This is primarily a function of extrinsic system properties, Z_k. Uncontrollable external factors, Y_i, alter both the intrinsic media properties,

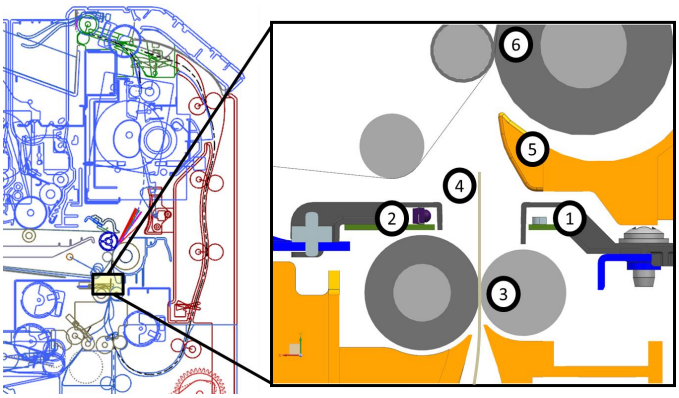


Fig. 3. A cross-section of the printer media path is depicted. The highlighted region contains an optical sensor consisting of an LED (1) and phototransistor (2) that measures the amount of infrared light transmitted by a sheet of media (4) as it is processed by the printer. Media fed by upstream feed rollers (3) passes through the sensor, beyond a media guide (5), and into a set of downstream feed rollers (6). Hardware (physical design of the media path) and firmware (system timings and relative velocities of the feed rollers) were tuned during development to enhance data orthogonality by controlling the position and shape of the media relative to the sensor in the spatial/temporal domain.

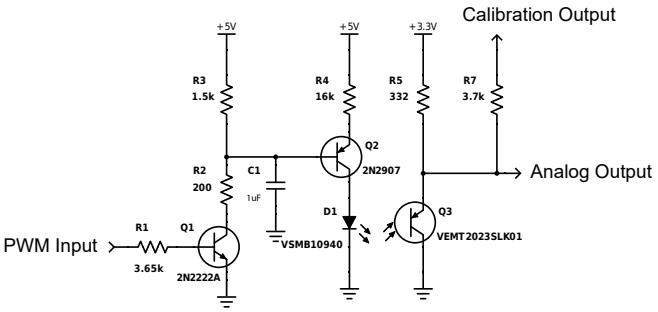


Fig. 4. The electrical design schematic for the optical sensor system is shown. The connection labeled "Analog Output" is the voltage signal measured by the analog-to-digital converter and used in the classification system. Nominal circuit values were selected to optimize the sensor gain, response, and sensitivity for a broad range of media types.

and extrinsic system properties, Z_k.

Figure 6 depicts how this variability manifests as boundary confusion. A broad set of standard office media possessing a range of intrinsic properties, φ_k, existing along a continuum were used to train and test the algorithm and are listed in Table II for reference. Corner cases (distinguishing light from card stock, for example) are easily distinguished. However, media properties exist along a continuum and variability from sheet to sheet and within a given sheet made the classification problem particularly challenging. There was a large amount of boundary confusion. This is especially true for the heavy class of media which significantly overlaps with both the normal and card stock classes.

For a classifier to be successful, it must decouple the relevant intrinsic media properties, φ_k, from the other confounding variables and generate a substantially orthogonal feature set. Media to media variability must be decoupled from the variability seen from sheet to sheet or within a given sheet. For the case of media classification, this was achieved

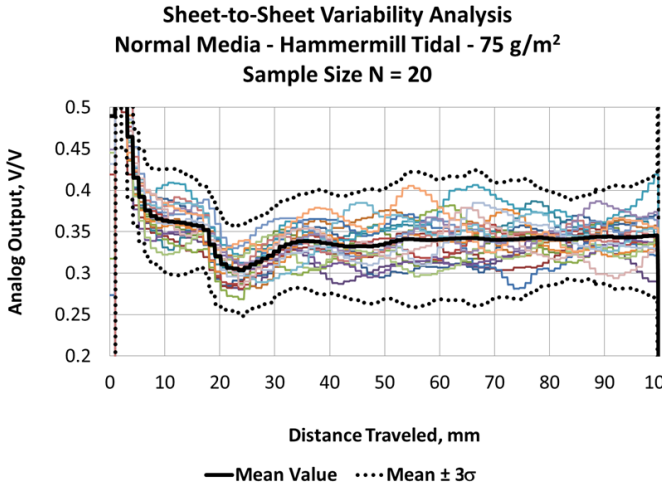


Fig. 5. Normalized analog sensor output for 20 separate sheets of a standard office paper are plotted. Data was collected for 100 millimeters of media travel. The population mean and 99.7 percent confidence bands for this given media are plotted for reference. Larger values on the y-axis correspond to an instant in time, or media leading edge position relative to the sensor, where less light was detected by the phototransistor. It is clear that there is a significant amount of noise in the measurement. This may be attributed to both intrinsic media properties, φ_k, extrinsic system properties, Z_k, and uncontrollable external factors,Y_i,.

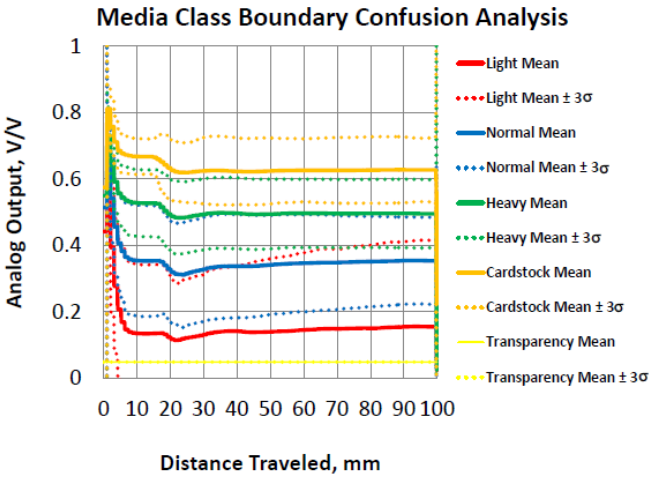


Fig. 6. Normalized analog sensor output for the mean and 99.7 percent confidence band of each class are plotted. The population for each class consists of 360 training samples from each media listed in Table II. A standard classification problem utilizing a traditional feature set would be intractable due to the continuous, overlapping nature of the data.

by tuning system hardware and control parameters to leverage the sensitivity of the measurement to uncontrollable external factors, Y_i, and extrinsic system properties, Z_k. Since the sensor output was a nonlinear function of φ_k, Z_k, and Y_i, it was possible to use the dynamic response of the system to help decouple these convoluted variables using the difference method described previously.

Concurrently developed printer control algorithms and sensor hardware were tuned during the development phase to generate a continuous data stream that could be deconstructed into several distinct zones of interest corresponding to the

sensors response to different dynamical processes. The resultant time series data was divided into 5 distinct zones of interest that corresponded to changes in the printer process that were designed to elicit a varied response from the sensor. In order to make the design more insensitive to printer-to-printer variation, four ideas were considered when designing the zone positions. First, a flag sensor (integrated into the paper feed control system) allowed accurate registration of the leading edge of the sheet, and the traverse distance was known from the paper feed drive encoders. Second, the zones are larger than strictly necessary for a single printer in order to accommodate variation around the population of printers (determined empirically from a number of different printers). It is important to be aware that performance can decrease if the buffer regions are too large as the data quality will decrease from the statistical measure being taken. Third, the features and zones are designed around bulk properties, as described in Figure 7, which are less sensitive to printer-to-printer variation. Finally, embedded firmware and system hardware were tuned during product development to generate subtle changes in media offset and shape relative to the emitter for each zone such that additional useful information may be extracted from the dataset.

This specific approach is summarized in Figure 7. For example, media in Zone 1 enters the sensor and obscures the photodetector. Prior to Zone 1, the photodetector is saturated and the signal is low. When the leading edge of the media directly obscures the direct path between the emitter and the photodetector, a minimal amount of light is transmitted and the signal is high. As the media continues downstream, a larger area of the in-process media is exposed to the emitter and additional diffusely scattered light reaches the phototector; the signal decreases. The output in Zone 1 is a strong function of media opacity and feed rate.

Further, media in Zone 3 is fed by two separate feed roller systems simultaneously. The relative velocity of the roller systems is precisely controlled by embedded firmware to elicit a specific media response. The shape of the bubble is strongly coupled to a specific intrinsic property (basis weight). Heavier media are stiffer and are less likely to buckle; the upstream feed rollers will slip. Lighter media will buckle and the position of the sheet relative to the sensor will change.

In this manner, the hardware and firmware within the system may be adjusted using expert domain knowledge to extract distinct information from the measurement based upon the dynamic response of the media to generated system inputs. This novel concurrent design approach allowed the photodetector to collect additional useful information that was strongly influenced by extrinsic system properties, Z_k.

Additionally, Zones 2, 3, and 4 extract similar information from the time series data. Each zone provides a distinct measure of media opacity that is a strong function of intrinsic media properties, φ_k. This provides the algorithm with a degree of redundancy and robustness against gross error.

Discretization of the analog data in this manner generated a richer feature set with some measurement redundancy. A small designed experiment was conducted to assess system performance and select the final feature set. Due to the information gained from the difference method previously described, inclusion of features from redundant zones yielded improved performance with minimal additional computing overhead. Features used for the machine learning algorithm are provided in Table I. Features $x_1, x_2, \ldots, x_5$ are extrinsic system properties and uncontrollable external factors that are provided by the printer systems embedded firmware to help stratify and decouple the training set. Features $x_6, x_7, \ldots, x_{18}$ contain an abundance of useful intrinsic media information, but are nonlinearly coupled to Z_k and Y_i. These features are calculated from the raw data and contain minimum, maximum and mean calculations (a measure of opacity) and range calculations (a measure of uniformity). Features x_{19}, x_{20}, x_{21} and x_{22} represent the previously described difference calculations that are used to separate φ_k information from the influence of Z_k and Y_i.

Constructing the feature set in this manner provided more useful information to the learning algorithm. The data set contains almost 7,000 examples divided into training, testing and validation sets. These examples represented data from four different process speeds and three different environments: hot/wet, lab ambient and cold/dry. The temperature and relative humidity were chosen as the corners of a Class B environment and represent the limits of the printer's operating space. A set of representative features gathered at a single process speed and environmental condition are shown in Figure 8. Feature 7 is predominantly a measure of the media opacity as the sheet passes through the sensor. Features 18, 19, and 20 are features that help decouple media opacity from other dynamic system interactions such as media sheet shape and position. The features in Figure 8 were selected to demonstrate that the difference method provided unique, pertinent information (Features 19 and 20) that allowed the learning algorithm to better classify media. This contention is supported by the different, distinct trends demonstrated by the plotted feature trends. However, all the features have significant boundary confusion, are not practical for use individually, but contribute to the overall classification performance.

After gathering a training set and developing an algorithm, the result was embedded as a set of decision polynomials used in a one vs. all classifier by the system firmware. This technique was effective in dealing with this type of classification problem and dramatically increased accuracy over a simple node mean comparison.

V. Implementation Performance

The results of the classification are given in Table II. The single node mean and the domain expert knowledge solutions are compared. The single node mean corresponds to the Zone 2 mean, or x_8, and was selected as the best single node classification system. The domain expert knowledge system was compared against this implementation. In the case of the domain expert knowledge, a number of feature sets using different order kernels were evaluated in a designed experiment to select the optimum group. A second order polynomial kernel with the features shown in Table I was selected. The cost function of the algorithm was modified to ensure media near

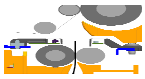
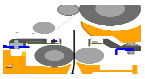
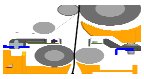
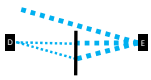
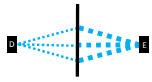
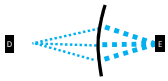
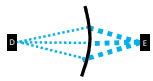
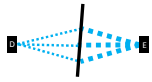
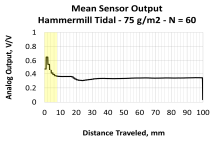
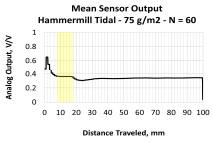
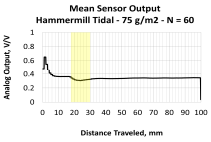
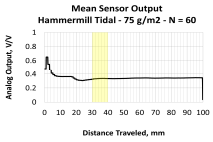
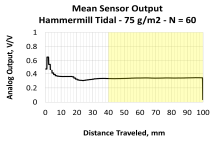
	Zone 1	Zone 2	Null	Zone 3	Zone 4
Distance Traveled	0 - 8 mm	8 - 18 mm	18 - 30 mm	30 - 40 mm	40 - 100 mm
Media Path Cross Section					
Physical Description	Media enters the sensor and obscures the emitter.	Media is vertical and midway between the emitter and detector.	Media bends and interacts with downstream guides.	Media enters a second nip. Shock is possible. A bubble is created.	A velocity differential creates tension and removes the bubble.
Media Sensor Pictogram					
Characteristic Sensor Output					
Explanation of Time Series	The detector is initially saturated. Sensor intensity drops drastically when the detector is first obscured and increases as previously undetectable light is diffusely scattered by media moving downstream.	Mean sensor intensity remains steady. Sheet position is well controlled near the center of the sensor.	Sensor intensity increases as media bends closer to the emitter. Output variability due to drastic fluctuations in sheet position are typical.	Some variability due to fluctuation in sheet position is typical. Once controlled by the second nip, media position begins to stabilize and higher sensor intensities are typical.	Mean sensor intensity remains steady. Sheet position is well controlled. Media is slightly angled and closer to the emitter. Higher intensities are typical.

Fig. 7. A simplified model of the response of the sensing system to the movement of media through the printer was developed. Observed measurement differences due to the dynamic interaction of the media with the sensor were leveraged to decouple confounded variables.

decision boundaries were classified in a manner that would have no negative impact on printer performance, as detailed in [23]. For this reason, "% Acceptable" is the key design metric for this system. This expert-prescribed cost function weighting resulted in one particular paper type (Canon GFR-070) having poorer "% Correct" than in the single node mean case. This particular error is due to the fact that media is not naturally categorical. The weighting method was designed to integrate into the printers existing control scheme with minimal system impact. The richer feature set provides the machine learning algorithm more flexibility to adjust decision surfaces such that printer performance is not compromised when boundary confusion occurs.

When the system is implemented using a single node mean method without domain expert knowledge, the resulting accuracy is 69% which filters to 93% acceptability when the expert-prescribed cost function weighting is applied. By comparison, the full implementation of domain expert knowledge including all features improved system performance to 85.38% correct with 99.95% acceptability. When the domain expert knowledge implementation did not contain the delta features (x_{19}, x_{20}, x_{21}, and x_{22}), and consisted solely of the minimum, maximum, and mean values from zones 1-3 ($x_6, x_7 - x_9, x_{11} - x_{13}, x_{15} - x_{17}$), classifier performance decreased to 81.46% correct with 98.93% acceptability. This increase in performance makes the case for value added by the domain expert knowledge method and the difference method previously described.

VI. Discussion

The methodology described has significant cost advantages over the traditional approach. These advantages stem from several fundamental aspects of single sensor design. This includes a reduction in hardware, associated non-recurring engineering expenses, and the savings associated with an inherent robustness of reducing the number of variables in the final system.

The case-study utilized a single optical sensor to extract multiple distinct pieces of information from a dynamic signal response. This means that only one sensor was required for the design. The development time was significantly shorter than that of a multi-sensor system. In addition to hardware cost reductions, the impact on the final product size was minimized. The resulting product had a reduced environmental footprint.

Reducing the part count also has some less obvious benefits. All sensors drift with time and have part-to-part variability. Ideally, products must account for this drift and variability. In a machine learning algorithm, many of the features being monitored and used are differences (deltas) or ratios. Relative calibration of different sensors becomes a performance-limiting design choice for a multi-sensor system. By utilizing only one sensor, changes in calibration factors are less complex to accommodate, both for the machine learning algorithm and for system design.

VII. Conclusions

A methodology for leveraging domain expert knowledge and temporal data for the design of an IoT system resulted in

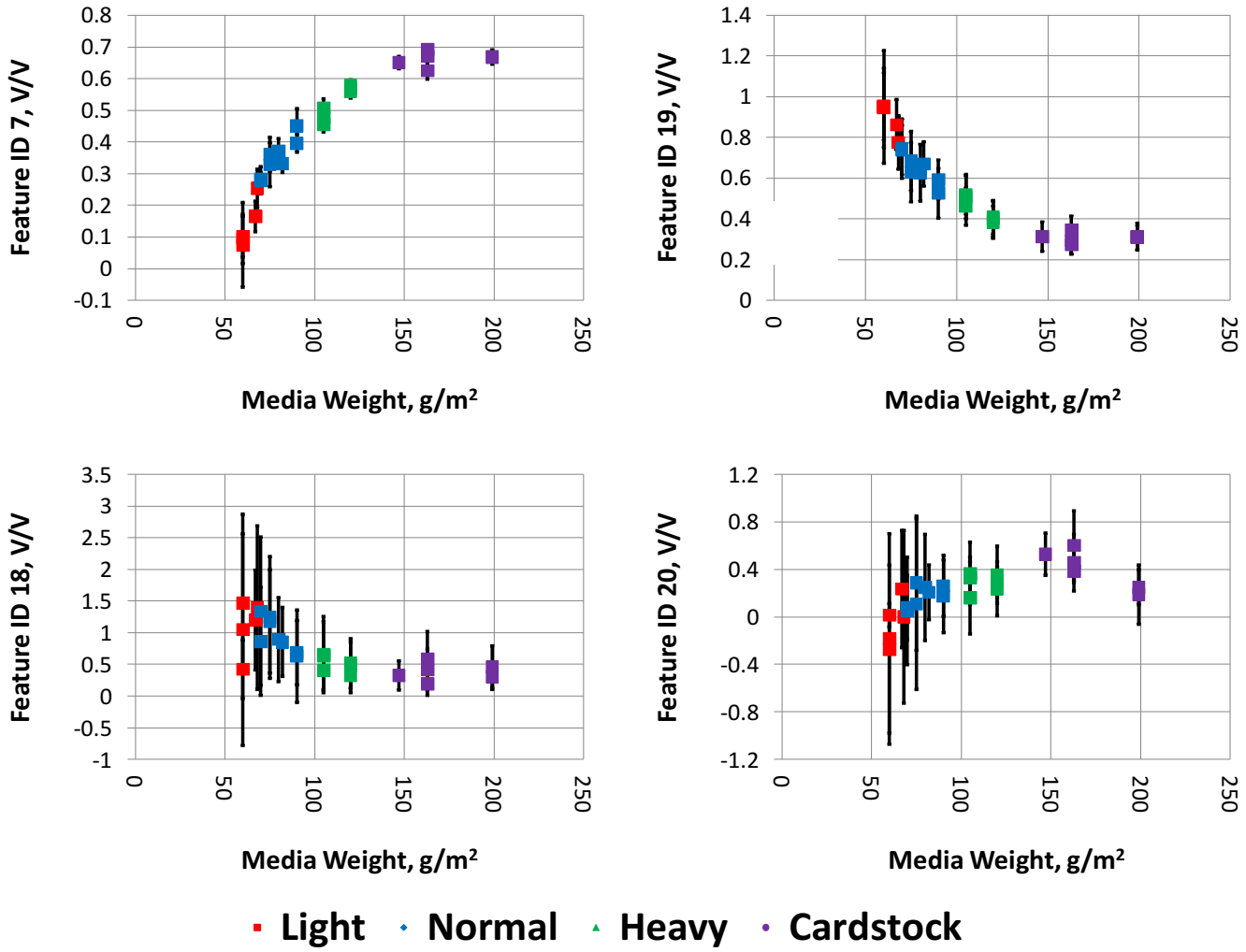


Fig. 8. Representative input features after scaling for an assortment of media types. While the features contain information that can do corner case separation, the features individually suffer from boundary confusion (significantly overlapping error bars between categories).

a highly robust and accurate system with reduced cost, size, and complexity over traditional approaches. This methodology was demonstrated in a case-study of a mass-produced electrophotographic printer in a system designed to classify media types. The proposed methodology increased classifier accuracy by 16% and classifier acceptability by 6.5% when compared with a more traditional method that did not leverage domain expert knowledge to enrich the dataset. The methodology used can be applied to sensor-integrated IoT devices seeking to benefit from performance enhancements associated with today's modern sensor technology while still meeting various market constraints.

REFERENCES

[1] N. Bajaj, G. T. C. Chiu, and J. P. Allebach, "Reduction of memory footprint and computation time for embedded support vector machine (SVM) by kernel expansion and consolidation," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2014, pp. 1–6.

[2] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun 1997, pp. 994–999.

[3] T. Mori, Y. Nejigane, M. Shimosaka, Y. Segawa, T. Harada, and T. Sato, "Online recognition and segmentation for time-series motion with HMM and conceptual relation of actions," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Aug 2005, pp. 3864–3870.

[4] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "An online algorithm for segmenting time series," in *Proceedings of the 2001 IEEE International Conference on Data Mining*, 2001, pp. 289–296.

[5] E. Fuchs, T. Gruber, J. Nitschke, and B. Sick, "Online segmentation of time series based on polynomial least-squares approximations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2232–2245, Sep. 2010.

[6] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Foundations of Data Organization and Algorithms*. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg, Oct. 1993, pp. 69–84.

[7] K.-P. Chan and A. W.-C. Fu, "Efficient time series matching by wavelets," in *Proceedings of the 15th International Conference on Data Engineering*, Mar 1999, pp. 126–133.

[8] I. Güler and E. D. Übeyli, "Multiclass support vector machines for EEG-signals classification," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 2, pp. 117–126, Mar. 2007.

[9] Y. Lei and M. J. Zuo, "Gear crack level identification based on weighted K nearest neighbor classification algorithm," *Mechanical Systems and Signal Processing*, vol. 23, no. 5, pp. 1535–1547, Jul. 2009.

[10] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27,

TABLE I
INPUT FEATURES USED BY THE MACHINE LEARNING ALGORITHM TO DETERMINE MEDIA CLASSIFICATION.

Feature ID	Description	Predominant Measure Type	Intuition
x_1	Process Speed (Discrete)	Extrinsic	Feed rate influences dynamic media bubble formation and data sampling rate
x_2	Temperature (Continuous)	External	Thermal expansion impacts roller diameter (feed rate)
x_3, x_4	Relative Humidity & Grains Moisture (Continuous)	External	Moisture content influences media stiffness (media bubble formation)
x_5	Input Source (Discrete)	Extrinsic	Input source influences media position and curl (media bubble formation)
x_6	Zone 1 Opacity (Max)	Intrinsic	Media opacity*, coupled with feed rate
x_7, x_8, x_9	Zone 2 Opacity (Min, Mean, & Max)	Intrinsic	Media opacity*
x_{10}	Zone 2 Uniformity (Range)	Intrinsic	Media uniformity*
x_{11}, x_{12}, x_{13}	Zone 3 Opacity (Min, Mean, & Max)	Intrinsic	Media opacity*, coupled with bubble formation
x_{14}	Zone 3 Uniformity (Range)	Intrinsic	Media uniformity*, coupled with bubble formation
x_{15}, x_{16}, x_{17}	Zone 4 Opacity (Min, Mean, & Max)	Intrinsic	Media opacity*, coupled with media offset
x_{18}	Zone 4 Uniformity (Range)	Intrinsic	Media uniformity*, coupled with media offset
x_{19}	$x_6 - x_7$ (Difference)	Decoupling	Decouples media opacity in Zone 1 from feed rate
x_{20}	$x_7 - x_{11}$ (Difference)	Decoupling	Decouples media opacity in Zone 2 from media bubble formation
x_{21}	$x_7 - x_{15}$ (Difference)	Decoupling	Decouples media opacity in Zone 4 from media offset
x_{22}	$x_{11} - x_{15}$ (Difference)	Decoupling	Helps decouple media bubble formation and media offset

*Function of media composition, thickness, roughness, etc.

TABLE II
THE RESULTS OF THE CLASSIFICATION ALGORITHM USING BOTH THE BEST SINGULAR-FEATURE THRESHOLD (x_8, THE ZONE 2 MEAN) AND THE FULL DOMAIN EXPERT KNOWLEDGE SUPPORTED MACHINE LEARNING ALGORITHM.

				Single Node Mean		Domain Knowledge	
Media ID	**Class**	**Description**	**Basis Weight**	**% Correct**	**% Acceptable**	**% Correct**	**% Acceptable**
1	**Light**	Boise X9	60 g/m^2	100.00%	100.00%	98.00%	99.00%
2		Clairmail Clairfontaine	60 g/m^2	99.62%	100.00%	100.00%	100.00%
3		Hp EcoFFICIENT	60 g/m^2	100.00%	100.00%	99.67%	100.00%
4		Ricoh My Paper	67 g/m^2	100.00%	100.00%	96.33%	100.00%
5		Canon GFR-070	68 g/m^2	66.15%	100.00%	4.67%	100.00%
6	**Normal**	Business (”C2”)	70 g/m^2	90.77%	90.77%	100.00%	100.00%
7		Sanyipaifuyinzhi	70 g/m^2	86.92%	86.92%	100.00%	100.00%
8		Husky Xerocopy	75 g/m^2	100.00%	100.00%	100.00%	100.00%
9		Hammermill Tidal	75 g/m^2	100.00%	100.00%	100.00%	100.00%
10		Datacopy	80 g/m^2	100.00%	100.00%	100.00%	100.00%
11		Premier ”J”	82 g/m^2	100.00%	100.00%	99.67%	100.00%
12		Domtar First Choice	90 g/m^2	51.54%	51.54%	95.67%	99.67%
13		Hammermill Laser Print	90 g/m^2	100.00%	100.00%	98.67%	100.00%
14	**Heavy**	Hammermill Laser Print	105 g/m^2	0.00%	100.00%	31.00%	100.00%
15		Boise Cascade Presentation	105 g/m^2	0.00%	100.00%	25.67%	100.00%
16		Via Satin Writing	105 g/m^2	0.00%	100.00%	35.33%	100.00%
17		Fine Color Copy Writing	120 g/m^2	0.00%	100.00%	72.67%	100.00%
18		Hammermill Laser Print	120 g/m^2	0.00%	100.00%	80.00%	100.00%
19		Boise Cascade Presentation	120 g/m^2	0.00%	100.00%	87.00%	100.00%
20	**Cardstock**	Exact Vellum Cover Bristol	147 g/m^2	100.00%	100.00%	100.00%	100.00%
21		Accent Opaque Digital	163 g/m^2	100.00%	100.00%	100.00%	100.00%
22		Hammermill Color Copy Cover	163 g/m^2	100.00%	100.00%	100.00%	100.00%
23		Springhill Index	163 g/m^2	100.00%	100.00%	95.67%	100.00%
24		Springhill Index	199 g/m^2	100.00%	100.00%	100.00%	100.00%
25		Exact Index	199 g/m^2	100.00%	100.00%	100.00%	100.00%
26	**Transparency**	Lexmark 70X7240	4.2 mil	0.00%	0.00%	100.00%	100.00%
		Total %		**69.02%**	**93.43%**	**85.38%**	**99.95%**

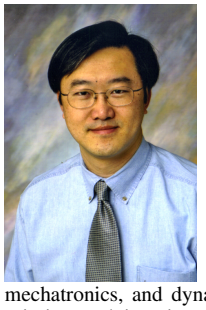
Jan. 1967.
[11] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast time series classification using numerosity reduction," in *The 23rd International Conference on Machine Learning*. New York, New York, USA: ACM Press, 2006, pp. 1033–1040.
[12] U. Orhan, M. Hekim, and M. Ozer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13 475–13 481, Sep. 2011.
[13] K. P. Murphy, *Machine learning: a probabilistic perspective*. Cambridge, Massachussetts, USA: The MIT Press, 2012.
[14] V. N. Vapnik, *The nature of statistical learning theory*. Springer, 1995.
[15] J. J. Rodríguez and C. J. Alonso, *Interval and dynamic time warping-based decision trees*. New York, New York, USA: ACM Press, 2004.
[16] C. Orsenigo and C. Vercellis, "Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification," *Pattern Recognition*, vol. 43, no. 11, pp. 3787–3794, Nov. 2010.
[17] J. S. Weaver, J. G. Bearss, and T. Camis, "Image forming devices and sensors configured to monitor media, and methods of forming an image upon media," U.S. Patent 6 157 793, Dec. 5, 2000.
[18] J. S. Weaver, "Capacitance and resistance monitor for image producing device," U.S. Patent 6 493 523, Dec. 10, 2002.
[19] M. Aoki, "Recording medium imaging apparatus for determining a type of a recording medium based on a surface image of a reference plate and a surface image of the recording medium," U.S. Patent 8 611 772, Sep. 16, 2013.
[20] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Dec. 17, 2013.
[21] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[22] J. A. K. Suykens, *Least squares support vector machines*. River Edge, NJ, USA: World Scientific, 2002.
[23] N. Bajaj, N. J. Murrell, J. G. Whitney, J. P. Allebach, and G. T. C. Chiu, "Expert-prescribed weighting for support vector machine classification," in *2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, July 2016, pp. 913–918.
[24] R. T. Bradley, J. L. Combs, N. J. Murrell, F. J. Palumbo, D. Steinberg, and J. A. G. Whitney, "Method of determining a media class in an imaging device using an optical translucence sensor," U.S. Patent 9 451 111, Sep. 20, 2016.

Niko Murrell received his BSME from Texas A&M University in 2001 and his MSME from the Georgia Institute of Technology in 2002 where he specialized in mechatronics and control systems. From 2003 to 2016, he worked for Lexmark in a variety of engineering roles including design, product development, supply chain, research, and systems engineering. He is currently employed by Ethicon as a staff electromechanical engineer and is supporting the digital surgery platform being developed by Verb Surgical (a joint venture between Johnson & Johnson's Ethicon and Alphabet's Verily Life Sciences). He is an inventor on 26 US patents and is a licensed professional mechanical engineer in the state of Kentucky.

Ryan Bradley earned his B.S. and M.S. degrees in Mechanical Engineering from the University of Kentucky, Lexington, KY, in 2014 and 2015, respectively. He launched his engineering career with the printer manufacturer, Lexmark International, in a variety of roles including research and development, sustainability, product engineering, and data science. Bradley is currently employed as a sustainability researcher, and he is currently pursuing a Ph.D. degree in Mechanical Engineering with the Institute of Sustainable Manufacturing (ISM), University of Kentucky. Through his research, he aspires to advance sustainable product design and manufacturing through the application of machine learning, life cycle assessment, life cycle costing, and alternative business models.

Julie Gordon Whitney earned her B.S. in Mechanical Engineering from Purdue University, her M.S. from Indiana State University and Ph.D. from the University of Cincinnati. Dr. Whitney has over 25 years of experience in Research and Development, and recently retired as a Senior Technical Staff Member from Lexmark International. She has joined the First Year Engineering program at the University of Kentucky as a Lecturer and enjoys traveling with her husband, Jon.

Nikhil Bajaj received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Purdue University, West Lafayette, IN, in 2008, 2011, and 2017, respectively. He is currently a postdoctoral research associate with the School of Mechanical Engineering, Purdue University, West Lafayette. He has held research assistant positions on several projects in the areas of advanced manufacturing, computational design, and heat transfer, and a summer research position with Alcatel-Lucent Bell Laboratories. His research interests include nonlinear dynamical and control systems, and the analysis and design of mechatronic systems, especially in the context of learning sensor systems.



George T.-C. Chiu received the B.S. degree from National Taiwan University, in 1985, and the M.S. and Ph.D. degrees from the University of California at Berkeley, in 1990 and 1994, respectively, all in mechanical engineering. He was with Hewlett-Packard, designing printers and multifunction devices. He is currently a Professor with the School of Mechanical Engineering with courtesy appointments in the School of Electrical and Computer Engineering, and the Department of Psychological Sciences, Purdue University. His current research interests are mechatronics, and dynamic systems and control with applications to digital printing and imaging systems, digital fabrications and functional printing, human motor control, and motion and vibration perception and control. He is a Fellow of ASME, and the Society for Imaging Science and Technology. He received The 2012 NSF Directors Collaboration Award, the 2010 IEEE TRANSACTIONS ON CONTROL SYSTEM TECHNOLOGY Outstanding Paper Award, the Purdue University College of Engineering Faculty Engagement/Service Excellence Award in 2010, and the Team Excellence Award in 2006. He was the Chair of the Executive Committee on the ASME Dynamic Systems and Control Division from 2013 to 2014.