Implementing Principal Component Analysis (PCA) from Scratch (July 2023)

Kaustuv Karki[1], Undergraduate, IOE, Nikhil Pradhan[2], Undergraduate, IOE

[1]Institute of Engineering Thapathali Campus, Tribhuvan University, Kathmandu, Nepal
[2]Institute of Engineering Thapathali Campus, Tribhuvan University, Kathmandu, Nepal

Corresponding author: Kaustuv Karki (e-mail: karkikaustuv@gmail.com), Nikhil Pradhan (e-mail: nikhilpradhan20b@gmail.com)

ABSTRACT Principal component analysis (PCA) is a widely used technique for dimensionality reduction and data visualization. It aims to find orthogonal vectors that capture the most variance in the data, and project the data onto a lower dimension to counter the problem created by high number of dimensions in a dataset. This report implements PCA from scratch using Python and NumPy and applies it to three datasets: a randomly generated dummy Dataset, Iris Dataset and Diabetes Prediction Dataset. After the data from the dataset were subjected to PCA and plots were generated to effectively visualize the transformed data. The dataset was subjected to our PCA implementation and the PCA from scikit-learn library and the results were compared. This report demonstrates the effectiveness of custom implementation of PCA and PCA from scikit-learn library for effective dimension reduction.

INDEX TERMS Covariance matrix, Dimensionality Reduction, Eigen Values, Eigen Vectors, Principal Component Analysis

I. INTRODUCTION

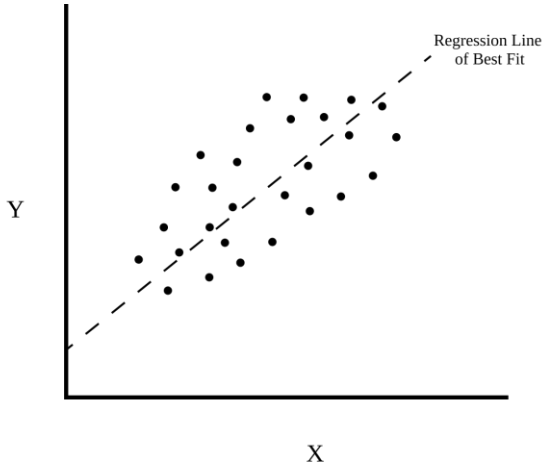
As the number of dimensions grows many problems arise in the dataset. The data becomes increasingly sparse meaning the datapoints are spread out very thinly making it hard to find any kind of pattern or relationship in the data and very high computational power is required for the high dimensional dataset. Also, there is a problem of visualizing the data in high dimensions. This problem is known as the curse of dimensionality. In the real world the collected data may have many features. Among these features there may be features whose values are irrelevant to the context of the dataset or are heavily correlated to other features present in the dataset. These features do not provide any valuable information or insights to the context of the dataset but only increase the complexity of the problem. So, it becomes a very important task to remove such irrelevant features and keep only the features which are relevant to the subject. And one of the methods to achieve this task is known as dimensionality reduction.

Principal component analysis (PCA) is a widely used technique for dimensionality reduction and data visualization. It aims to reduce the dimension of the dataset while preserving as much information as possible. It aims to find a set of orthogonal vectors, called principal components, that capture the most variance in the data. The orthogonal vectors are in such order that the first principal component captures the most variance while each subsequent principal component captures the next highest variance in the dataset. The principal components are orthogonal to each other signifying that they are not correlated with each other.

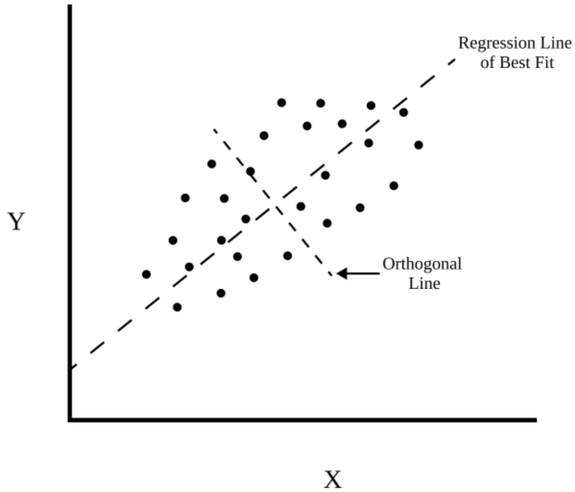
II. METHODOLOGY

A. BRIEF THEORY

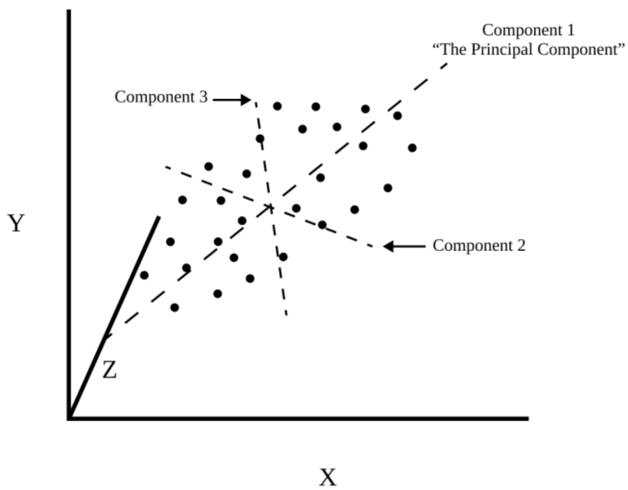
The main concept of PCA is removing any redundant dimensions and keeping the dimensions with the highest variance. PCA selects the principal components that capture all the major variations across the dataset encompassing most of the information present in the dataset.



Suppose we have a dataset with features X and Y then we can fit a regression line onto the dataset. Drawing an orthogonal line to the best fit line we get,



We can see that the data vary mostly along the best fit line. Now, we can project the points onto our new axis and get our new x-axis and y-axis. We can keep drawing lines perpendicular to both lines to get more axis.



This new axis will be our Principal Components. PC1 is generally used to denote the component that captures the most variation, PC2 the next and so on.

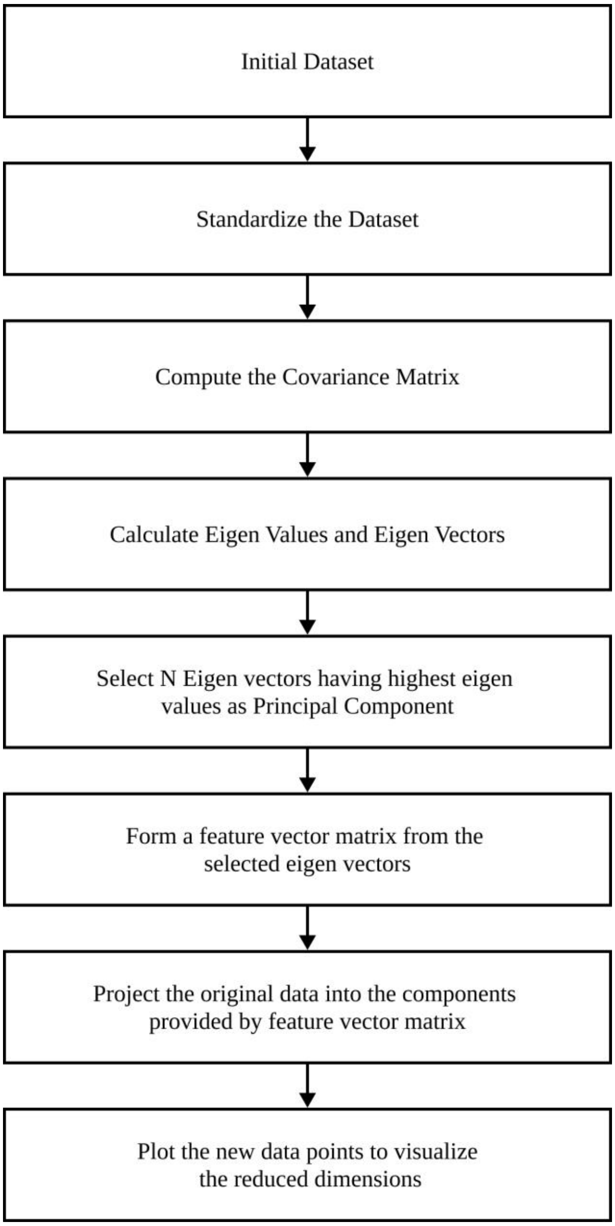
B. WORKING PRINCIPLE

Initially PCA was performed on a random (20 X 2) matrix was generated from a standard normal distribution. Thus, generated matrix had no correlation so PCA could not be applied. Random (2 X 2) matrix was then generated from a uniform distribution. This matrix was then multiplied with the original matrix to generate a new (20 X 2) matrix. Newly generated data was then plotted which resulted in a correlated data point. Standardization was not required as the data from standard normal distribution had 0 mean and multiplying with the (2 X 2) matrix is scaling. Covariance matrix was calculated this, and its value was checked. The required criteria of diagonal element high and off diagonal element nearly 0 was not achieved. Eigenvalues and eigenvectors were extracted from this covariance matrix. Initially both principal components were taken, and a change of basis vector was performed. The variance along Principal Component 1 was very high compared to Principal Component 2 so PC2 was ignored and PC1 was taken. PCA was again using the scikit-learn library and then compared with PCA from scratch. Thus, PCA of random data was completed.

PCA for iris data set firstly involved creating a new data frame and loading the iris data set into the data frame. Features were separated from the class and a matrix of features was created. Calculation of mean and standard deviation showed that the dataset needed to be standardized. Standardization was performed using StandardScaler from scikit-learn. Thus, the dataset had mean 0 and standard deviation of 1. Covariance matrix of the standardized matrix was calculated which did not satisfy our required properties. New basis vectors for the dataset were calculated by eigen decomposition of covariance matrix. Eigenvectors were used to perform changes of basis and projected features were obtained. PCA was again done using decomposition from scikit-learn and compared with PCA from scratch. Thus, PCA of random data was completed.

Diabetes prediction dataset was chosen as our next candidate for PCA. Dataset consisted of 8 features and class. Categorial classes were converted into numerical form. Some of the features were NaN (Not a Number) which was filled with mean value using SimpleImputer. Features was separated from the class and matrix was created. Calculation of mean and standard deviation showed that the dataset needed to be standardized. Standardization was performed. Thus, the dataset had mean 0 and standard deviation of 1. Covariance matrix of the standardized matrix was calculated which did not satisfy our required properties. New basis vectors for the dataset were calculated by eigen decomposition of covariance matrix. Eigenvectors were used to perform change of basis and projected features was obtained. PCA was again done using scikit-learn library and compared with PCA from scratch. Newly obtained datapoints were the plotted to analyze the result. Thus, PCA of random data was completed.

C. SYSTEM BLOCK DIAGRAM



1) CREATE A MATRIX OF FEATURE VECTORS

Initially features are selected from which the target is interpreted. These features allow the model to make predictions on new data to assign classes.

2) STANDARDIZE THE DATA

Before starting the PCA process, the feature vectors must be standardized so that they have zero mean and standard deviation of 1. This ensures that all the features have the same scale so that PCA gives equal importance to each feature preventing features with larger scale to disproportionately influence the result. Units are also removed during standardization making interpretation of principal components more meaningful.

3) COVARIANCE MATRIX COMPUTATION

After standardization of features, covariance matrix of the standardized features is calculated. Covariance matrix shows the linear relationship between the features. For PCA decomposition, the diagonal elements of the covariance matrix i.e., Variance of a feature should be maximum, and the non-diagonal elements of the covariance matrix must be minimum. Covariance matrix allows PCA to find the direction of maximum variance. (eqn no 3)

4) CALCULATION OF EIGEN VALUES AND EIGEN VECTORS

The covariance matrix is used of eigen decomposition to get the eigenvalues and eigenvectors. Eigenvalues represent the amount of variance captured by each principal component and eigenvectors represent the direction of each principal component. Each eigenvector is new basis vectors that are used to project the original data into new coordinates.

5) SELECTION OF PRINCIPAL COMPONENT

Eigen values are used to calculate the importance of each principal component. If n number of feature is required, then eigen values are arranged in ascending order and eigenvectors with top n eigenvalue is selected. This ensures that maximum amount of variance is captured thus retaining essential information.

6) FORMATION OF FEATURE MATRIX

Selected eigenvectors are used to form a feature new feature matrix. The feature matrix acts as a new basis for the data, and the new features are orthogonal to each other.

7) PROJECT ORIGINAL DATA INTO THE PRINCIPAL COMPONENTS

Change of basis is performed on the original dataset to project them onto new coordinates. This results in new dataset with reduced number of features. (Change of bias refer to eqn no 4)

8) VISUALIZATION

New data is then plotted to visualize the changes study the result.

B. MAJOR MATHEMATICAL FORMULAS

1) VARIANCE

Variance is used to measure how scattered the data are from a mean value. It is calculated by.

$$Var(x) = \frac{\sum_{i=0}^{n}(x_i - \bar{x})}{n-1} \quad (1)$$

Where:
x_i = Individual data points
$\bar{x}$ = Mean of all the value
n = Number of data points in the given data.

2) COVARIANCE

Covariance is the measure of variance between variables. It describes how variable change to together. Covariance is calculated as

$$Cov(x,y) = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad (2)$$

Where:
x_i = Individual data points of feature x
$\bar{x}$ = Mean of all the value of feature x
y_i = Individual data points of feature y
$\bar{y}$ = Mean of all the value of feature y
n = Number of data points in the given data.

Covariance gives an idea about the trend of the data. Positive covariance indicates that with the increase of x, y also increases whereas negative covariance indicates that with increase in x, y decreases. Zero covariance indicates that there is no relationship between x and y.

3) COVARIANCE MATRIX

Covariance Matrix is a square matrix with diagonal elements representing variance and non-diagonal elements representing covariance. Generally, Covariance matrix is represented as

$$\begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{bmatrix}$$

Where:
Var(x) = Variance of x
Var(y) = Variance of y
Cov(x,y) = Covariance between x and y

Covariance Matrix is calculated as

$$S_x = \frac{1}{n-1} XX^T \quad (3)$$

Where:
S_x = Square symmetric (n*n) matrix
X = n-dimensional vector of feature x

4) CHANGE OF BASIS

Basis of a vector space is the set of linearly independent vectors that span all the vector space. For e.g. Consider a 2-dimensional space with unit vector $\hat{i}$ along x axis and $\hat{j}$ along y axis. These unit vectors can then be scaled so that it can span all the possible points in the 2-dimensional space. In PCA, change on basis plays a vital role to as we map the original data into new one by changing the basis vector. Changing the basis does not change the data only its representation is changed. Changing the basis only projects the data vectors on the basis vectors.

Change of basis is achieved by

$$Y = \text{PX} \quad (4)$$

Where:
Y = Data points obtained after linear transformation
P = Basis vectors used for linear transformation
X = Original data points

3) EIGEN VALUES AND EIGEN VECTORS

Eigen v
Let A be an $n \times n$ matrix, then

An eigen vector is a non-zero vector v such that

$$Av = \lambda v \quad (5)$$

For some scalar λ.

An eigen value is a scalar λ such that

$$Av = \lambda v \quad (5)$$

Has some non-trivial solution.

If $Av = \lambda v$ and $v \neq 0$, then λ is eigenvalue for v, and v is the eigen vector for λ.

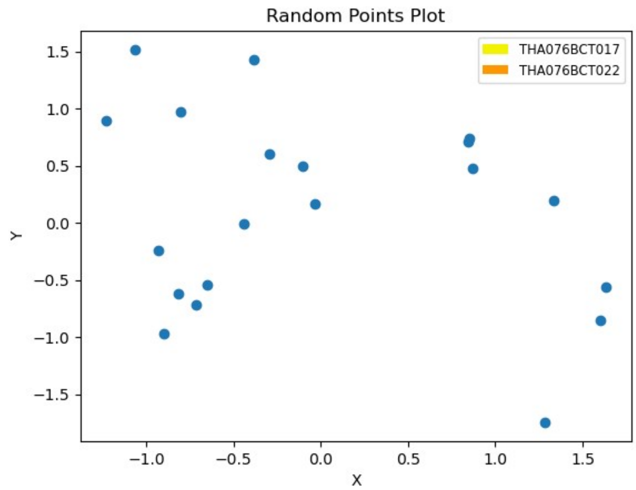
All eigen vectors of a symmetric matrix are perpendicular to each other.

C. INSTRUMENTATION

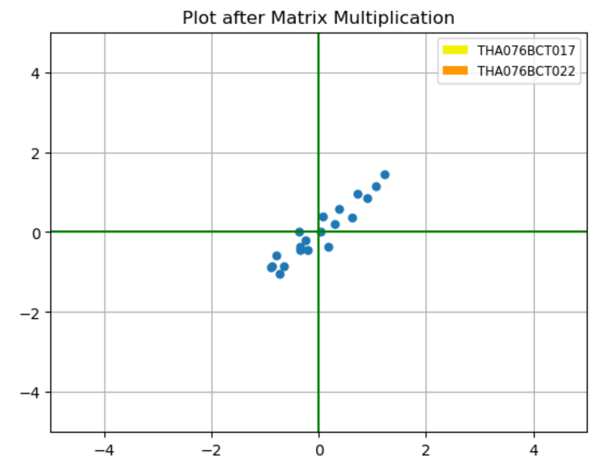
In the implementation of Principal Component Analysis (PCA) using Python, the following libraries and functions were utilized. Numpy, a powerful library for numerical computing, was employed for vector and matrix operations, providing efficient computation capabilities. Pandas, a popular data manipulation library, was used for storing and handling data in the form of dataframes. Scikit-learn, a comprehensive machine learning library, contributed the StandardScaler class for scaling the data, ensuring that all features have similar ranges. The np.linalg.eig function from Numpy was utilized to calculate the eigenvalues and eigenvectors, crucial components of PCA. Lastly, the PCA class from scikit-learn was employed to compare the results against the hand-coded implementation, facilitating a convenient and validated approach to PCA analysis.

III. RESULTS

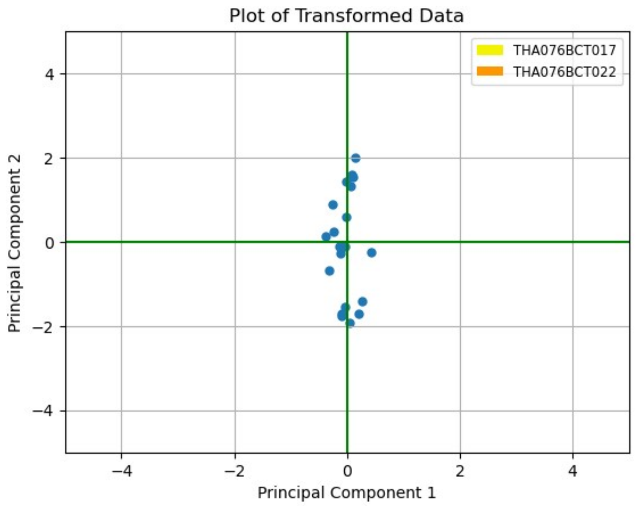
A. RANDOM DATASET



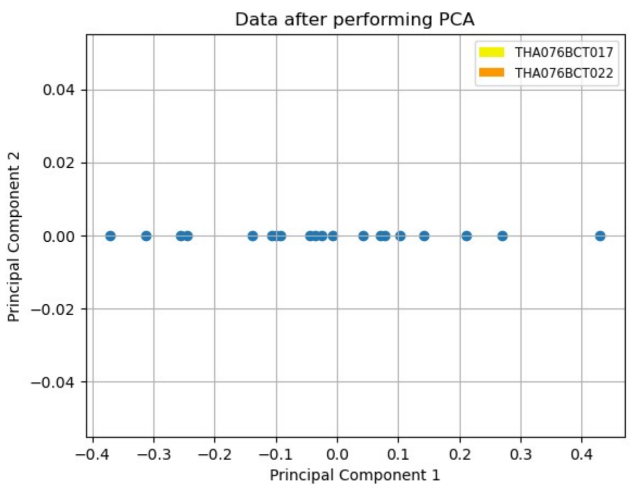
The plot shows the datapoint obtained from standard normal distribution. Since the data is extracted randomly from standard normal distribution, datapoints show no correlation thus PCA cannot be applied. The values generated form a standard normal distribution generates a mean of zero and standard deviation close to 1.



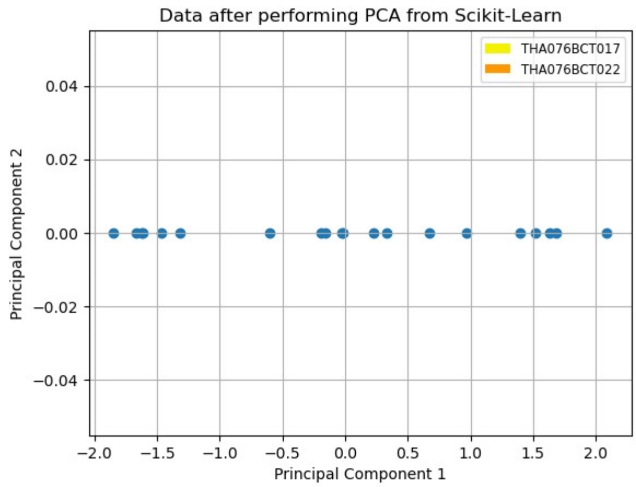
Multiplication with (2 X 2) matrix from uniform distribution resulted in the above plot. It is seen that the datapoints are correlated. This is because multiplication with the (2 X 2) matrix introduces a linear transformation upon the data. Because the transformation involves a linear combination of the original variables leading to a change in their relationship, as a result the transformed data points become correlated.



The plot was obtained by performing a change of basis vector and data point was projected upon new coordinated system defined by the principal components generated by applying PCA. The Principal Component 1 is as x-axis and Principal Component 2 is at y-axis.



The plot above demonstrated the reduction of the original two-dimensional data into a single dimension using PCA. In this plot the data are varied only in a single dimension x-axis using principal component 1 and the y-axis value which is principal component 2 remains zero.



The above plot was obtained by using the PCA provided by scikit-learn library to reduce the dimension from two-dimensions to a single dimension.

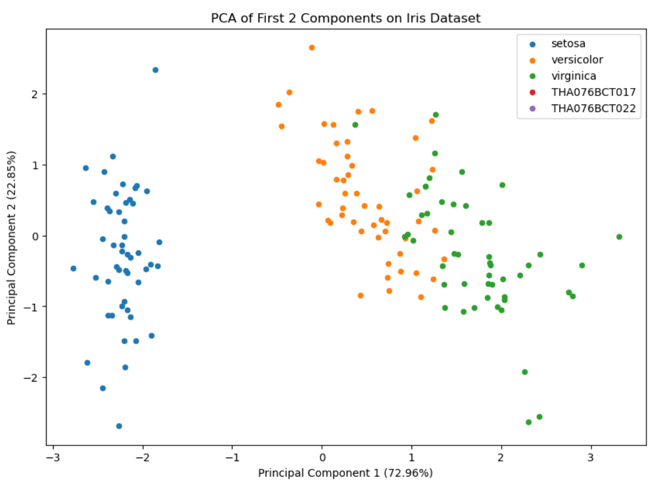
B. IRIS DATASET

Like the random generated dataset PCA was also applied to the Iris Dataset. The results obtained from applying the PCA on the Iris Dataset are discussed below.

TABLE I
EXPLAINED VARIANCE FOR IRIS DATASET

Principal Component	Explained Variance	Approx. (in %)
PC1	0.7296	72.96
PC2	0.2285	22.85
PC3	0.03669	3.66
PC4	0.00517	0.51

Table I gives the values of variance captured by every principal component after applying PCA.



The above plot new data points of Iris Dataset after undergoing PCA. This 2-D plot depicts PC1 which covers 72.96% of the variation and PC2 which covers 22.85% of the variation. With only two features of the four, setosa can be seen separated from the other two flowers but there still lies some confusion between versicolor and virginica.

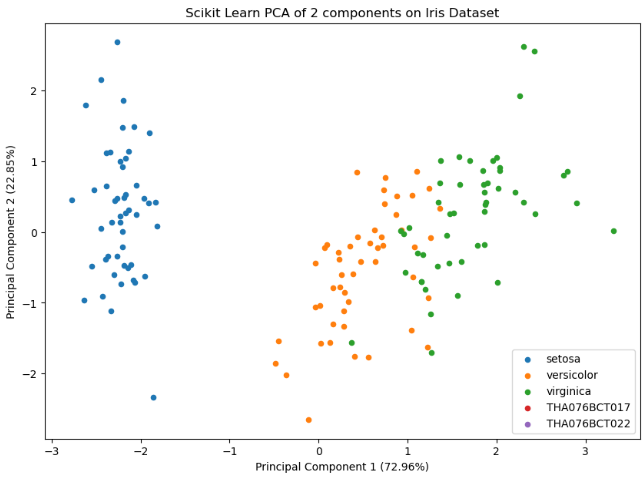


Figure shows the top two PCA accomplished using the decomposition function from scikit-learn. We have observed that the eigenvalues and eigenvectors result in the same value using both methods i.e., applying PCA from scratch and PCA from scikit-learn. The only difference observed was the direction of eigenvectors which was introduced due to a sign difference. This sign difference does not cause any issue for PCA. But for the plot it caused the plot to flip along the y-axis. This occurred due to some difference in mathematical implementation.

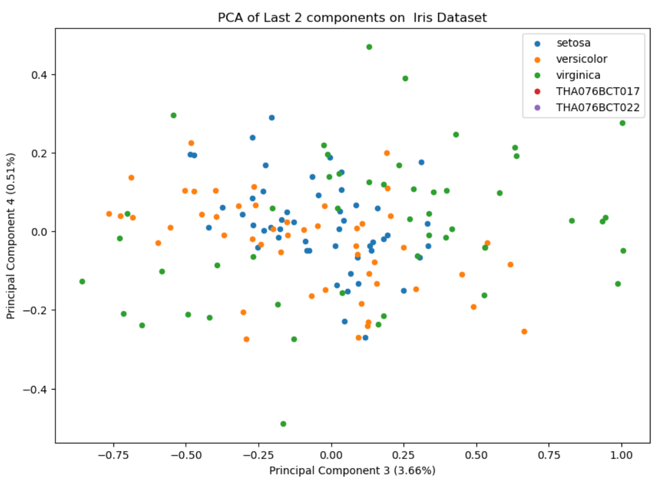


Figure shows the PC3 which captures 3.66% of the variation and PC4 which captures 0.51% of the variation. Separating the flowers was not possible from the available data points as most of the essential information about the original data was lost because the principal components could not retain most of the information from the dataset.

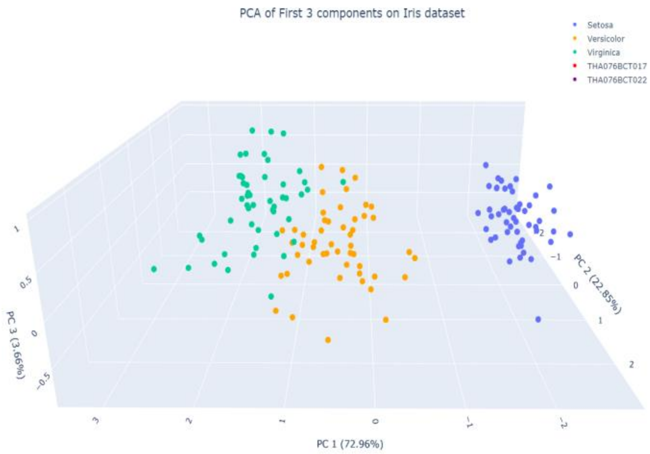


Figure shows 3D plot consisting of PC1, PC2 and PC3. This combination of Principal Component captures the largest amount of variation in the dataset while being easily visualizable. Flowers can be seen to be easily separatable as crucial information of the original dataset was retained.

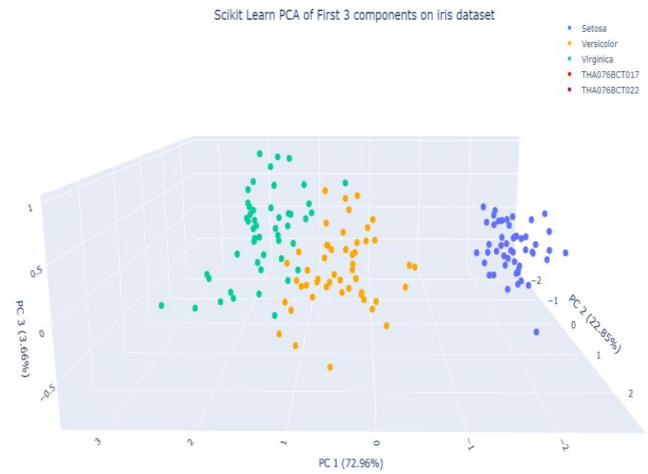


Figure shows top three PC implemented from decomposition using the PCA from scikit-learn library.

Figure shows the 3D plot of PC with least amount of variation captured. As crucial information is lost, separation of the flowers cannot be done.

C. DIABETES DATASET

Finally, PCA was applied to the Diabetes Detection Dataset. The results obtained from applying the PCA on the Diabetes Detection Dataset are discussed below.

TABLE II
EXPLAINED VARIANCE FOR DIABETES PREDICTION DATASET

Principal Component	Explained Variance	Approx. (in %)
PC1	0.2150	21.5
PC2	0.1380	13.8
PC3	0.1322	13.22
PC4	0.1196	11.96
PC5	0.1099	10.99
PC6	0.1056	10.56
PC7	0.1041	10.41
PC8	0.0752	7.52

Table II gives the values of variance captured by every principal component after applying PCA on the diabetes detection dataset.

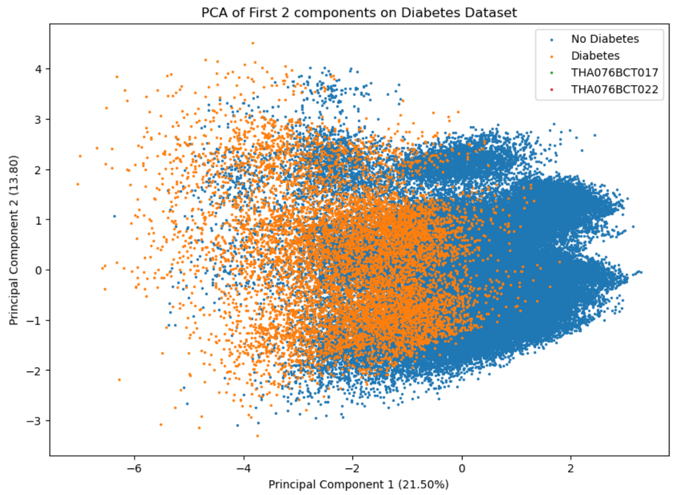


Figure shows the top two principal components for diabetes prediction. Since the original dataset had 8 dimensions so two principal components are not enough to show the difference so not many insights can be gained from this figure.

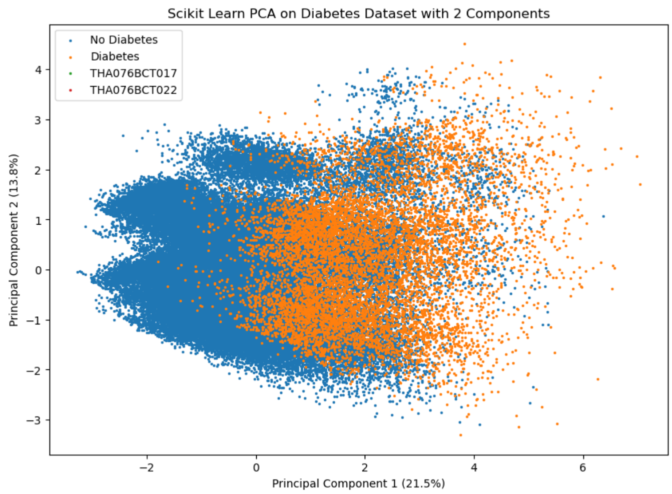


Figure shows top two PC from decomposition from scikit learn. Because of the mathematical implementation of PCA in scikit-learn library and custom one we can see that the plots are flipped on the x-axis.

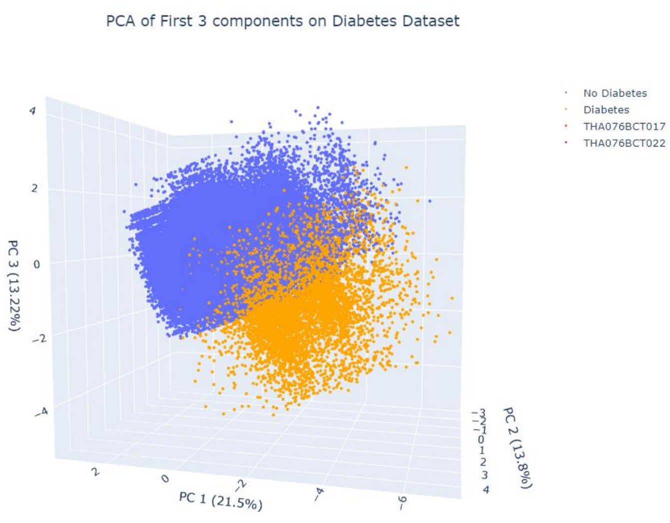


Figure shows 3-d plot of top three PC for diabetes prediction.

Figure shows principal components obtained from scikit-learn that capture the least amount of variance in diabetes prediction dataset.

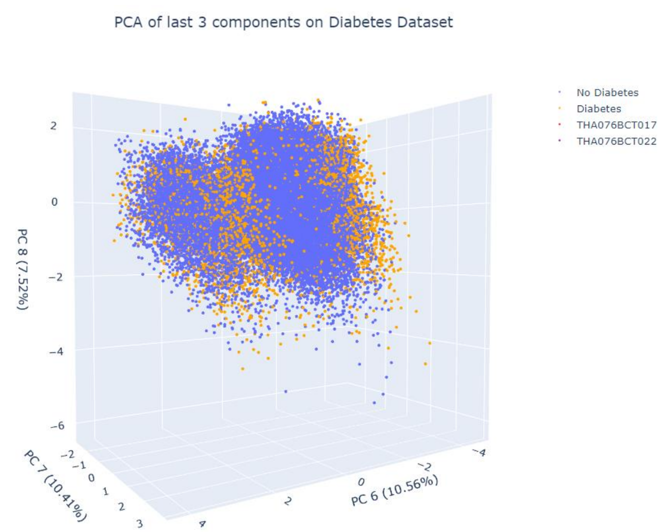


Figure shows principal components that capture the least amount of variance in diabetes prediction dataset.

IV. DISCUSSION AND ANALYSIS

Use either SI (MKS) or CGS as primary units. (SI units are strongly encouraged.) English units may be used as secondary units (in parentheses). This applies to papers in data storage. For example, write “15 Gb/cm^2 (100 Gb/in^2).” An exception is when English units are used as identifiers in trade, such as “3½-in disk drive.” Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity in an equation.

The SI unit for magnetic field strength H is A/m. However, if you wish to use units of T, either refer to magnetic flux density B or magnetic field strength symbolized as $\mu_0 H$. Use the center dot to separate compound units, e.g., “A·m^2.”

V. CONCLUSION

The word “data” is plural, not singular. The subscript for the permeability of vacuum μ_0 is zero, not a lowercase letter “o.” The term for residual magnetization is “remanence”; the adjective is “remanent”; do not write “remnance” or “remnant.” Use the word “micrometer” instead of “micron.” A graph within a graph is an “inset,” not an “insert.” The word “alternatively” is preferred to the word “alternately” (unless you really mean something that alternates). Use the word “whereas” instead of “while” (unless you are referring to simultaneous events). Do not use the word “essentially” to mean “approximately” or “effectively.” Do not use the word “issue” as a euphemism for “problem.” When compositions are not specified, separate chemical symbols by en-dashes; for example, “NiMn” indicates the intermetallic compound $Ni_{0.5}Mn_{0.5}$ whereas “Ni–Mn” indicates an alloy of some composition Ni_xMn_{1-x}.

Be aware of the different meanings of the homophones "affect" (usually a verb) and "effect" (usually a noun), "complement" and "compliment," "discreet" and "discrete," "principal" (e.g., "principal investigator") and "principle"

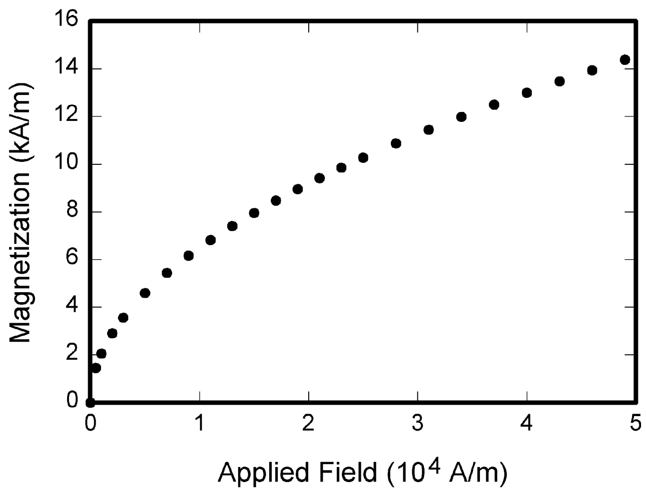


FIGURE 1. **Magnetization as a function of applied field. Note that "Fig." is abbreviated. There is a period after the figure number, followed by two spaces. It is good practice to explain the significance of the figure in the caption.**

VI. REFERENCES

D. SIZING OF GRAPHICS

Most charts, graphs, and tables are one column wide (3.5 inches / 88 millimeters / 21 picas) or page wide (7.16 inches / 181 millimeters / 43 picas). The maximum depth a graphic can be is 8.5 inches (216 millimeters / 54 picas). When choosing the depth of a graphic, please allow space for a caption. Figures can be sized between column and page widths if the author chooses, however it is recommended that figures are not sized less than column width unless when necessary.

There is currently one publication with column measurements that do not coincide with those listed above. Proceedings of the IEEE has a column measurement of 3.25 inches (82.5 millimeters / 19.5 picas).

The final printed size of author photographs is exactly 1 inch wide by 1.25 inches tall (25.4 millimeters x 31.75 millimeters / 6 picas x 7.5 picas). Author photos printed in editorials measure 1.59 inches wide by 2 inches tall (40 millimeters x 50 millimeters / 9.5 picas x 12 picas).

E. RESOLUTION

The proper resolution of your figures will depend on the type of figure it is as defined in the "Types of Figures" section. Author photographs, color, and grayscale figures should be at least 300dpi. Line art, including tables should be a minimum of 600dpi.

F. VECTOR ART

In order to preserve the figures' integrity across multiple computer platforms, we accept files in the following formats: .EPS/.PDF/.PS. All fonts must be embedded or text converted to outlines in order to achieve the best-quality results.

G. COLOR SPACE

The term color space refers to the entire sum of colors that can be represented within the said medium. For our purposes, the three main color spaces are Grayscale, RGB (red/green/blue) and CMYK (cyan/magenta/yellow/black). RGB is generally used with on-screen graphics, whereas CMYK is used for printing purposes.

All color figures should be generated in RGB or CMYK color space. Grayscale images should be submitted in Grayscale color space. Line art may be provided in grayscale OR bitmap colorspace. Note that "bitmap colorspace" and "bitmap file format" are not the same thing. When bitmap color space is selected, .TIF/.TIFF/.PNG are the recommended file formats.

H. ACCEPTED FONTS WITHIN FIGURES

When preparing your graphics IEEE suggests that you use of one of the following Open Type fonts: Times New Roman, Helvetica, Arial, Cambria, and Symbol. If you are supplying EPS, PS, or PDF files all fonts must be embedded. Some fonts may only be native to your operating system; without the fonts embedded, parts of the graphic may be distorted or missing.

A safe option when finalizing your figures is to strip out the fonts before you save the files, creating "outline" type. This converts fonts to artwork what will appear uniformly on any screen.

I. USING LABELS WITHIN FIGURES

1) FIGURE AXIS LABELS

Figure axis labels are often a source of confusion. Use words rather than symbols. As an example, write the quantity "Magnetization," or "Magnetization *M*," not just "*M*." Put units in parentheses. Do not label axes only with units. As in Fig. 1, for example, write "Magnetization (A/m)" or "Magnetization ($A \cdot m^{-1}$)," not just "A/m." Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)," not "Temperature/K."

Multipliers can be especially confusing. Write "Magnetization (kA/m)" or "Magnetization (10^3 A/m)." Do not write "Magnetization (A/m) × 1000" because the reader would not know whether the top axis label in Fig. 1 meant 16000 A/m or 0.016 A/m. Figure labels should be legible, approximately 8 to 10 point type.

2) SUBFIGURE LABELS IN MULTIPART FIGURES AND TABLES

Multipart figures should be combined and labeled before final submission. Labels should appear centered below each subfigure in 8 point Times New Roman font in the format of (a) (b) (c).

J. FILE NAMING

Figures (line artwork or photographs) should be named starting with the first 5 letters of the author's last name. The next characters in the filename should be the number that represents the sequential location of this image in your article. For example, in author "Anderson's" paper, the first three figures would be named ander1.tif, ander2.tif, and ander3.ps.

Tables should contain only the body of the table (not the caption) and should be named similarly to figures, except that '.t' is inserted in-between the author's name and the table number. For example, author Anderson's first three tables would be named ander.t1.tif, ander.t2.ps, ander.t3.eps.

Author photographs should be named using the first five characters of the pictured author's last name. For example, four author photographs for a paper may be named: oppen.ps, moshc.tif, chen.eps, and duran.pdf.

If two authors or more have the same last name, their first initial(s) can be substituted for the fifth, fourth, third... letters of their surname until the degree where there is differentiation. For example, two authors Michael and Monica Oppenheimer's photos would be named oppmi.tif, and oppmo.eps.

K. REFERENCING A FIGURE OR TABLE WITHIN YOUR PAPER

When referencing your figures and tables within your paper, use the abbreviation "Fig." even at the beginning of a sentence. Do not abbreviate "Table." Tables should be numbered with Roman Numerals.

L. CHECKING YOUR FIGURES: THE IEEE GRAPHICS ANALYZER

The IEEE Graphics Analyzer enables authors to pre-screen their graphics for compliance with IEEE Access standards before submission. The online tool, located at http://graphicsqc.ieee.org/, allows authors to upload their graphics in order to check that each file is the correct file format, resolution, size and colorspace; that no fonts are missing or corrupt; that figures are not compiled in layers or have transparency, and that they are named according to the IEEE Access naming convention. At the end of this automated process, authors are provided with a detailed report on each graphic within the web applet, as well as by email.

For more information on using the Graphics Analyzer or any other graphics related topic, contact the IEEE Graphics Help Desk by e-mail at graphics@ieee.org.

M. SUBMITTING YOUR GRAPHICS

Because IEEE will do the final formatting of your paper, you do not need to position figures and tables at the top and bottom of each column. In fact, all figures, figure captions, and tables can be placed at the end of your paper. In addition to, or even in lieu of submitting figures within your final manuscript, figures should be submitted individually, separate from the manuscript in one of the file formats listed above in section VI-J. Place figure captions below the figures; place table titles above the tables. Please do not include captions as part of the figures, or put them in "text boxes" linked to the figures. Also, do not place borders around the outside of your figures.

N. COLOR PROCESSING / PRINTING IN IEEE JOURNALS

All IEEE Transactions, Journals, and Letters allow an author to publish color figures on IEEE *Xplore*® at no charge, and automatically convert them to grayscale for print versions. In most journals, figures and tables may alternatively be printed in color if an author chooses to do so. Please note that this service comes at an extra expense to the author. If you intend to have print color graphics, include a note with your final paper indicating which figures or tables you would like to be handled that way, and stating that you are willing to pay the additional fee.

VII. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in American English is without an "e" after the "g." Use the singular heading even if you have many acknowledgments. Avoid expressions such as "One of us (S.B.A.) would like to thank ... ." Instead, write "F. A. Author thanks ... ." In most cases, sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

REFERENCES AND FOOTNOTES

A. REFERENCES

References need not be cited in text. When they are, they appear on the line, in square brackets, inside the punctuation. Multiple references are each numbered with separate brackets. When citing a section in a book, please give the relevant page numbers. In text, refer simply to the reference number. Do not use "Ref." or "reference" except at the beginning of a sentence: "Reference [3] shows ... ." Please do not use automatic endnotes in *Word*, rather, type the reference list at the end of the paper using the "References" style.

Reference numbers are set flush left and form a column of their own, hanging out beyond the body of the reference. The reference numbers are on the line, enclosed in square brackets. In all references, the given name of the author or editor is abbreviated to the initial only and precedes the last name. Use them all; use *et al.* only if names are not given. Use commas around Jr., Sr., and III in names. Abbreviate conference titles. When citing IEEE transactions, provide the issue number, page range, volume number, year, and/or month if available. When referencing a patent, provide the day and the month of issue, or application. References may not include all information; please obtain and include relevant information. Do not combine references. There must be only one reference with each number. If there is a URL included with the print reference, it can be included at the end of the reference.

Other than books, capitalize only the first word in a paper title, except for proper nouns and element symbols. For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation See the end of this document for formats and examples of common references. For a complete discussion of references and their formats, see the IEEE style manual at www.ieee.org/authortools.

A. FOOTNOTES

Number footnotes separately in superscripts (Insert| Footnote).[1] Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

VIII. SUBMITTING YOUR PAPER FOR REVIEW

A. REVIEW STAGE USING WORD 6.0 OR HIGHER

If you want to submit your file with one column electronically, please do the following:

--First, click on the View menu and choose Print Layout.

--Second, place your cursor in the first paragraph. Go to the Format menu, choose Columns, choose one column Layout, and choose "apply to whole document" from the dropdown menu.

--Third, click and drag the right margin bar to just over 4 inches in width.

The graphics will stay in the "second" column, but you can drag them to the first column. Make the graphic wider to push out any text that may try to fill in next to the graphic.

B. FINAL STAGE USING WORD 6.0

When you submit your final version (after your paper has been accepted), print it in two-column format, including figures and tables. You must also send your final manuscript on a disk, via e-mail, or through a Web manuscript submission system as directed by the society contact. You may use *Zip* for large files, or compress files using *Compress, Pkzip, Stuffit,* or *Gzip.*

Also, send a sheet of paper or PDF with complete contact information for all authors. Include full mailing addresses, telephone numbers, fax numbers, and e-mail addresses. This information will be used to send each author a complimentary copy of the journal in which the paper appears. In addition, designate one author as the "corresponding author." This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only.

C. REVIEW STAGE USING SCHOLARONE® MANUSCRIPTS

Contributions to the Transactions, Journals, and Letters may be submitted electronically on IEEE's on-line manuscript submission and peer-review system, ScholarOne® Manuscripts. You can get a listing of the publications that participate in ScholarOne at http://www.ieee.org/publications_standards/publications/authors/authors_submission.html First check if you have an existing account. If there is none, please create a new account. After logging in, go to your Author Center and click "Submit First Draft of a New Manuscript."

[1]It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, try to integrate the footnote information into the text.

Along with other information, you will be asked to select the subject from a pull-down list. Depending on the journal, there are various steps to the submission process; you must complete all steps for a complete submission. At the end of each step you must click "Save and Continue"; just uploading the paper is not sufficient. After the last step, you should see a confirmation that the submission is complete. You should also receive an e-mail confirmation. For inquiries regarding the submission of your paper on ScholarOne Manuscripts, please contact oprs-support@ieee.org or call +1 732 465 5861.

ScholarOne Manuscripts will accept files for review in various formats. Please check the guidelines of the specific journal for which you plan to submit.

You will be asked to file an electronic copyright form immediately upon completing the submission process (authors are responsible for obtaining any security clearances). Failure to submit the electronic copyright could result in publishing delays later. You will also have the opportunity to designate your article as "open access" if you agree to pay the IEEE open access fee.

D. FINAL STAGE USING SCHOLARONE MANUSCRIPTS

Upon acceptance, you will receive an email with specific instructions regarding the submission of your final files. To avoid any delays in publication, please be sure to follow these instructions. Most journals require that final submissions be uploaded through ScholarOne Manuscripts, although some may still accept final submissions via email. Final submissions should include source files of your accepted manuscript, high quality graphic files, and a formatted pdf file. If you have any questions regarding the final submission process, please contact the administrative contact for the journal.

In addition to this, upload a file with complete contact information for all authors. Include full mailing addresses, telephone numbers, fax numbers, and e-mail addresses. Designate the author who submitted the manuscript on ScholarOne Manuscripts as the "corresponding author." This is the only author to whom proofs of the paper will be sent.

E. COPYRIGHT FORM

Authors must submit an electronic IEEE Copyright Form (eCF) upon submitting their final manuscript files. You can access the eCF system through your manuscript submission system or through the Author Gateway. You are responsible for obtaining any necessary approvals and/or security clearances. For additional information on intellectual property rights, visit the IEEE Intellectual Property Rights department web page at http://www.ieee.org/publications_standards/publications/rights/index.html.

IX. IEEE PUBLISHING POLICY

The general IEEE policy requires that authors should only submit original work that has neither appeared elsewhere for publication, nor is under review for another refereed publication. The submitting author must disclose all prior publication(s) and current submissions when submitting a manuscript. Do not publish "preliminary" data or results. The Department strongly discourages courtesy authorship; it is the obligation of the authors to cite only relevant prior work.

The IEEE Access Department does not publish conference records or proceedings, but can publish articles related to conferences that have undergone rigorous peer review. Minimally, two reviews are required for every article submitted for peer review.

X. PUBLICATION PRINCIPLES

The two types of contents of that are published are; 1) peer-reviewed and 2) archival. The Transactions and Journals Department publishes scholarly articles of archival value as well as tutorial expositions and critical reviews of classical subjects and topics of current interest.

Authors should consider the following points:

1) Technical papers submitted for publication must advance the state of knowledge and must cite relevant prior work.
2) The length of a submitted paper should be commensurate with the importance, or appropriate to the complexity, of the work. For example, an obvious extension of previously published work might not be appropriate for publication or might be adequately treated in just a few pages.
3) Authors must convince both peer reviewers and the editors of the scientific and technical merit of a paper; the standards of proof are higher when extraordinary or unexpected results are reported.
4) Because replication is required for scientific progress, papers submitted for publication must provide sufficient information to allow readers to perform similar experiments or calculations and use the reported results. Although not everything need be disclosed, a paper must contain new, useable, and fully described information. For example, a specimen's chemical composition need not be reported if the main purpose of a paper is to introduce a new measurement technique. Authors should expect to be challenged by reviewers if the results are not supported by adequate data and critical details.
5) Papers that describe ongoing work or announce the latest technical achievement, which are suitable for presentation at a professional conference, may not be appropriate for publication.

REFERENCES

Basic format for books:

J. K. Author, "Title of chapter in the book," in *Title of His Published Book, x*th ed. City of Publisher, (only U.S. State), Country: Abbrev. of Publisher, year, ch. *x*, sec. *x*, pp. *xxx–xxx.*

Examples:

submitting author is responsible for obtaining agreement of all coauthors and any consent required from employers or sponsors before submitting an article. The IEEE Access

[1] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.
[2] W.-K. Chen, *Linear Networks and Systems.* Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

Basic format for periodicals:

J. K. Author, "Name of paper," *Abbrev. Title of Periodical*, vol. *x, no*. *x,* pp. *xxx-xxx,* Abbrev. Month, year, DOI. 10.1109.*XXX*.123456.

Examples:

[3] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, 10.1109/TED.2016.2628402.
[4] E. P. Wigner, "Theory of traveling-wave optical laser," *Phys. Rev*., vol. 134, pp. A635–A646, Dec. 1965.
[5] E. H. Miller, "A note on reflector arrays," *IEEE Trans. Antennas Propagat*., to be published.

Basic format for reports:

J. K. Author, "Title of report," Abbrev. Name of Co., City of Co., Abbrev. State, Country, Rep. *xxx*, year.

Examples:

[6] E. E. Reber, R. L. Michell, and C. J. Carter, "Oxygen absorption in the earth's atmosphere," Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.
[7] J. H. Davis and J. R. Cogdell, "Calibration program for the 16-foot antenna," Elect. Eng. Res. Lab., Univ. Texas, Austin, TX, USA, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

Basic format for handbooks:

Name of Manual/Handbook, x ed., Abbrev. Name of Co., City of Co., Abbrev. State, Country, year, pp. *xxx-xxx.*

Examples:

[8] *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, USA, 1985, pp. 44–60.
[9] *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, USA, 1989.

Basic format for books (when available online):

J. K. Author, "Title of chapter in the book," in *Title of Published Book*, *x*th ed. City of Publisher, State, Country: Abbrev. of Publisher, year, ch.*x*, sec. *x*, pp. *xxx–xxx*. [Online]. Available: http://www.web.com

Examples:

[10] G. O. Young, "Synthetic structure of industrial plastics," in Plastics, vol. 3, Polymers of Hexadromicon, J. Peters, Ed., 2nd ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15-64. [Online]. Available: http://www.bookref.com.
[11] *The Founders' Constitution*, Philip B. Kurland and Ralph Lerner, eds., Chicago, IL, USA: Univ. Chicago Press, 1987. [Online]. Available: http://press-pubs.uchicago.edu/founders/
[12] The Terahertz Wave eBook. ZOmega Terahertz Corp., 2014. [Online]. Available: http://dl.z-thz.com/eBook/zomega_ebook_pdf_1206_sr.pdf. Accessed on: May 19, 2014.
[13] Philip B. Kurland and Ralph Lerner, eds., *The Founders' Constitution.* Chicago, IL, USA: Univ. of Chicago Press, 1987, Accessed on: Feb. 28, 2010,

[Online] Available: http://press-pubs.uchicago.edu/founders/

Basic format for journals (when available online):

J. K. Author, "Name of paper," *Abbrev. Title of Periodical*, vol. *x*, no. *x*, pp. *xxx-xxx*, Abbrev. Month, year. Accessed on: Month, Day, year, DOI: 10.1109.*XXX*.123456, [Online].

Examples:

[14] J. S. Turner, "New directions in communications," *IEEE J. Sel. Areas Commun*., vol. 13, no. 1, pp. 11-23, Jan. 1995.

[15] W. P. Risk, G. S. Kino, and H. J. Shaw, "Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle," *Opt. Lett.*, vol. 11, no. 2, pp. 115–117, Feb. 1986.

[16] P. Kopyt *et al., "*Electric properties of graphene-based conductive layers from DC up to terahertz range," *IEEE THz Sci. Technol.,* to be published. DOI: 10.1109/TTHZ.2016.2544142.

Basic format for papers presented at conferences (when available online):

J.K. Author. (year, month). Title. presented at abbrev. conference title. [Type of Medium]. Available: site/path/file

Example:

[17] PROCESS Corporation, Boston, MA, USA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annual Meeting. [Online]. Available: http://home.process.com/Intranets/wp2.htp

Basic format for reports and handbooks (when available online):

J. K. Author. "Title of report," Company. City, State, Country. Rep. no., (optional: vol./issue), Date. [Online] Available: site/path/file

Examples:

[18] R. J. Hijmans and J. van Etten, "Raster: Geographic analysis and modeling with raster data," R Package Version 2.0-12, Jan. 12, 2012. [Online]. Available: http://CRAN.R-project.org/package=raster

[19] Teralyzer. Lytera UG, Kirchhain, Germany [Online]. Available: http://www.lytera.de/Terahertz_THz_Spectroscopy.php?id=home, Accessed on: Jun. 5, 2014

Basic format for computer programs and electronic documents (when available online):

Legislative body. Number of Congress, Session. (year, month day). *Number of bill or resolution*, *Title*. [Type of medium]. Available: site/path/file

NOTE: ISO recommends that capitalization follow the accepted practice for the language or script in which the information is given.

Example:

[20] U.S. House. 102nd Congress, 1st Session. (1991, Jan. 11). *H. Con. Res. 1, Sense of the Congress on Approval of Military Action*. [Online]. Available: LEXIS Library: GENFED File: BILLS

Basic format for patents (when available online):

Name of the invention, by inventor's name. (year, month day). Patent Number [Type of medium]. Available: site/path/file

Example:

[21] Musical toothbrush with mirror, by L.M.R. Brooks. (1992, May 19). Patent D 326 189 [Online]. Available: NEXIS Library: LEXPAT File: DES

Basic format for conference proceedings (published):

J. K. Author, "Title of paper," in *Abbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), Country, year, pp. *xxxxxx*.

Example:

[22] D. B. Payne and J. R. Stern, "Wavelength-switched passively coupled single-mode optical network," in *Proc. IOOC-ECOC,* Boston, MA, USA, 1985, pp. 585–590.

Example for papers presented at conferences (unpublished):

[23] D. Ebehard and E. Voges, "Digital single sideband detection for interferometric sensors," presented at the *2nd Int. Conf. Optical Fiber Sensors,* Stuttgart, Germany, Jan. 2-5, 1984.

Basic format for patents:

J. K. Author, "Title of patent," U.S. Patent *x xxx xxx*, Abbrev. Month, day, year.

Example:

[24] G. Brandli and M. Dick, "Alternating current fed power supply," U.S. Patent 4 084 217, Nov. 4, 1978.

Basic format for theses (M.S.) and dissertations (Ph.D.):

a) J. K. Author, "Title of thesis," M.S. thesis, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

b) J. K. Author, "Title of dissertation," Ph.D. dissertation, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

Examples:

[25] J. O. Williams, "Narrow-band analyzer," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.

[26] N. Kawasaki, "Parametric study of thermal and chemical nonequilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

Basic format for the most common types of unpublished references:

a) J. K. Author, private communication, Abbrev. Month, year.

b) J. K. Author, "Title of paper," unpublished.

c) J. K. Author, "Title of paper," to be published.

Examples:

[27] A. Harrison, private communication, May 1995.

[28] B. Smith, "An approach to graphs of linear forms," unpublished.

[29] A. Brahms, "Representation error for real numbers in binary computer arithmetic," IEEE Computer Group Repository, Paper R-67-85.

Basic formats for standards:

a) *Title of Standard*, Standard number, date.

b) *Title of Standard*, Standard number, Corporate author, location, date.

Examples:

[30] IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969.

[31] Letter Symbols for Quantities, ANSI Standard Y10.5-1968.

Article number in reference examples:

[32] R. Fardel, M. Nagel, F. Nuesch, T. Lippert, and A. Wokaun, “Fabrication of organic light emitting diode pixels by laser-assisted forward transfer,” *Appl. Phys. Lett.*, vol. 91, no. 6, Aug. 2007, Art. no. 061103.

[33] J. Zhang and N. Tansu, “Optical gain and laser characteristics of InGaN quantum wells on ternary InGaN substrates,” *IEEE Photon. J.*, vol. 5, no. 2, Apr. 2013, Art. no. 2600111.

Example when using et al.:

[34] S. Azodolmolky *et al.*, Experimental demonstration of an impairment aware network planning and operation tool for transparent/translucent optical networks,” *J. Lightw. Technol.*, vol. 29, no. 4, pp. 439–448, Sep. 2011.

KAUSTUV KARKI (M’76–SM’81–F’87) and all authors may include biographies. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author’s major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author’s last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

The third paragraph begins with the author’s title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, it should be of good quality, and professional-looking. Following are two examples of an author’s biography.

NIKHIL PRADHAN was born in Greenwich Village, New York, NY, USA in 1977. He received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008.

From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth*, *Moon*, *Planets*, and holds two patents.

Dr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.