An Anatomy of Data Visualization

Abhishek Kaushik[†] and Sudhanshu Naithani[††],

Kiel University of Applied Sciences[†] Kurukshetra University [††]

Summary
As data is being generated each and every time in the world, the importance of data mining and visualization will always be on increase. Mining helps to extract significant insight from large volume of data. After that we need to present that data in such a way so that it can be understood by everyone and for that visualization is used. Most common way to visualize data is chart and table. Visualization is playing important role in decision making process for industry. Visualization makes better utilization of human eyes to assist his brain so that datasets can be analyzed and visual presentation can be prepared. Visualization and Data Mining works as complement for each other. Here in this paper we present anatomy of Visualization process.

Key words:
Information Visualization, Scientific Visualization, Decision Making, Graph, Chart, Xmdv tool.

1. Introduction

In simple worlds Visualization is a process to form a picture in order to make it easily imaginable and understandable for other people. With Visualization, process of Data Mining and Human Computer Interaction provides better results for visual data analysis. Initially visualization was of two types - Information Visualization and Scientific Visualization. Scientific Visualization used to work for scientific data with spatial component while Information Visualization used to work for abstract and non-spatial data [12]. Presently visualization is facing problems like mapping, dimensionality, and design tradeoff [13]. Visualization helps to understand patterns, trends and relationship between different components in a dataset. In words of David McCandless [32] (author, data journalist, and information designer) :- "By visualizing information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you're lost in Information, an information map is kind of useful."
Figure 1 shows the general steps in the process of Visualization. For visualization data is collected from all the available sources. Then possible aggregate meaning is generated. After that data is analyzed. After it, graphical interpretation of analyzed data takes place. And at last step user interacts with graphical interpretation.
Here are basics to generate best possible visualization [27] for any given data:-

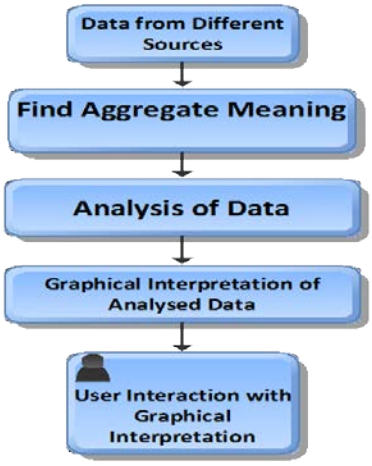

Fig.1. General steps in the process of Visualization.

a) Try to understand size and cardinality of the data given.
b) Determine kind of information which is to communicate.
c) Process visual information according to targeted audience.
d) Use the visual portraying best and easiest form of given data for audience.

2. Classes of Data Visualization Techniques

The most common classes [2] of data visualization techniques are:

a) Describing Data
b) Viewing Relationship
c) Picturing Data (Icons, Glyphs, Color Coding)
d) Temporal Visualization
e) Spatial Visualization
f) Spatio-Temporal Visualization

Class (a) tells about the dataset. Class (b) describes relationship between observations and between variables. Class (c) maps data items into easily recognizable shapes. Class (d) describes visualization of temporal data which changes over time. Line graph is most suitable in this case. Class (e) describes spatial datasets which come from various domains that relate data to a certain landscape [2].

Manuscript received February 5, 2016
Manuscript revised February 20, 2016

Map is utilized for this type of data. Class (f) has both kinds of properties i.e. spatial and temporal such as analysis of biomedical data.

3. State of the Art

Ming C. Hao [14] proposed Visual Analysis of Multi-Attribute Data Using Pixel Matrix Displays, which is especially useful in case of detailed information. Adam Perer [15] proposed that tight integration of statistical and visualization techniques could speed up insight development. Zhao Kaidi [4] presented a new algorithm for 4D data visualization. Martin Wattenberg [6] gave method of arc diagram for visualization. Ming C. Hoa [3] proposed two new techniques for visual analytics which are cell based visual time series and visual content query. Doantam Phan [9] presented a method for generating flow maps using hierarchical clustering. Mohammad Daradkeh [24, 25] designed new InfoVis tool to support informed decision-making under uncertainty and risk through Interactive Visualisation. Kristine Amari [28] presented techniques and tools for recovering and analyzing data from volatile memory. Jarkko Venna [30] introduced NeRV (neighbor retrieval visualizer) to produces an optimal visualization by minimizing the cost. Sandro Boccuzzo [16] addressed software comprehension by a combination of visualization and audio. Pak Chung Wong [17] presented visualize association rule for text mining. Svetlana Mansmann [20] proposed an explorative framework for OLAP data to analyzing data cubes of virtually arbitrary complexity. Ji Soo Yi [31] developed InfoVis tool to improve decision quality of nursing home choice.

4. Methods of Visualization (with examples)

4.1 Arc Diagram

Arc diagram is usually used to visualize complex data within string such as text, music, compile code. In structure of string there are repetitions of sub-string most of the time, which is a good thing as point of view of visualization because these repetitions can be used as prediction units for the visualization process. For example in any given article there will be repetition of words and phrases. Martin Wattenberg [6] described arc diagram visualization to process string by using pattern matching algorithm to find repeated substring and further representing them visually as translucent arcs. Most significant utilization of arc diagram is in the field of music which is to reveal structure in compositions of music [6]. Other utilization fields for arc diagrams are web pages, compiled codes, and nucleotide sequence from DNA etc. In future other pattern matching algorithms can be used for arc diagrams to get unknown insights of this method.

4.2 Flow Maps

As name suggests flow maps show the flow of any process i.e. how particular process is flowing. For example when people migrate from one country to another, a flow map can show this very easily. Doantam Phan [9] presented a method to generate flow map using hierarchical clustering which is inspired by graph layout algorithm. In a hand written flow map intelligent distortion of positions, merging of edges that share destination and intelligent edge routing are most common characteristics [9]. To achieve intelligent distortion, Doantam used layout adjustment algorithm. For merging edges and intelligent edge routing, hierarchical clustering is used. To implement this system an algorithm was used with following steps:-

a) Layout Adjustment
b) Primary Hierarchical Clustering
c) Rooted Hierarchical Clustering
d) Spatial Layout
e) Edge Routing
f) Multiple –Layer Issues (when there are multiple layers is in the system)

4.3 Graph Analytics

It is common and interesting topic of visualization and analytics. The main aim of graph analytics research is to meet real life challenges. It follows technology-application pair i.e. success is measured by application and not by algorithmic criteria [11]. In the applications it always turn lesson learned into lesson applied. The main applications (real life challenges) of graph analytics are listed following [11]:-

a) Electric-Power-Grid Analytics
b) Social-Network and Citation Analytics
c) Text and Document Analytics
d) Knowledge Domain Analytics

4.4 Voronoi Treemaps

For visualization of attributed hierarchical data Treemaps are best method. Treemaps normally has problem of rectangular shapes limitation, which is removed by using Voronoi Treemaps. It also enables arbitrary shape visualization [37]. Michael Balzer [37] presented Treemaps based on the subdivision in arbitrary polygons which eliminate rectangle limitation because of recursive pattern. In this system following steps are repeated again and again:-

a) According to top hierarchy level polygonal subdivision of display area is created.
b) output is a set of polygons representing the nodes of the top hierarchy level

4.5 Geometric Projection

It is a technique used for multidimensional multivariate visualization. It can map Cartesian plane as well as arbitrary space [13]. This technique is good to detect outliers and handle large datasets.

Scatterplot is used to show joint variation of 2 data items at x-y axes of Cartesian coordinates. It supports grouping. In case of 3 or more measures a matrix named scatter plot matrix is produced which is a series of scatter plots to display possible pairing of measures that are assigned to visualization [27].

Parallel Coordinates technique is used where attributes are represented by parallel vertical axes linearly scaled within their data range [13]. Coordinates also utilized to study correlations among attributes by locating points of intersection [13]. Here limited space is available for a parallel axis. It has mainly two types i.e. circular and hierarchical.

4.6 Pixel-Oriented Technique

It is also used to visualize multivariate data where an attribute is represented by colored pixel. In n-dimensional dataset to represent a data item n colored pixel will be used.

Recursive Pattern can influence data arrangements by using generic recursive process. It is a query independent.
Pixel Bar Chart does not aggregate data values but presents them directly. These are derived from regular bar charts. Multi-pixel bar charts are used for high dimensional data [13].

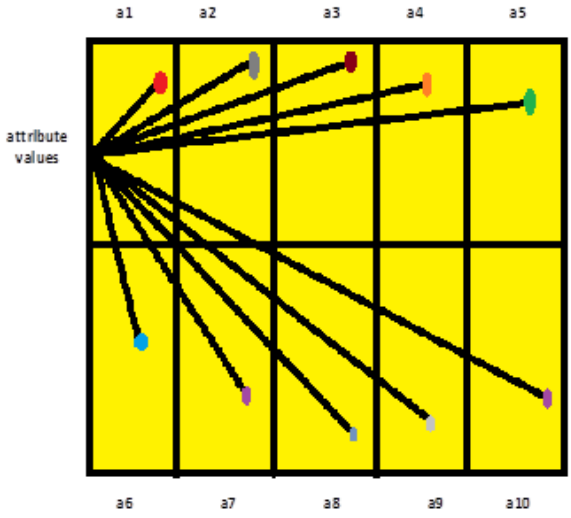

Fig. 2 Pixel Visualization if 10 Dimensional Data (a is attribute)

4.7 Hierarchical Display

These techniques are mainly concerned about hierarchical data where data space is sub divided first and then sub spaces are presented in hierarchical way.
Dimensional Stacking which is also known as general logic diagrams is a technique which is result of modification in hierarchical axis. It separates the data space into 2D stacked subspaces [13].

Treemap partitions the screen into several regions on behalf of value of attribute by using hierarchical partitioning.

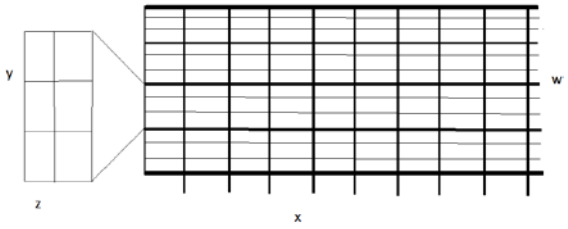

Fig. 3 Dimensional Stacking.

4.8 Iconography

It maps multidimensional data item to an icon and also known as icon-based techniques. The visual features vary depending on the data attribute values [13].

Chernoff Faces which is most popular technique of Iconography can visualize data items in a limit. It maps dimensional positions of a face and its properties like mouth, eyes and nose etc [13].

Star Glyph is one of the many variants of glyph family and is most widely used. Here star glyphs are used to present data items. It is not suitable when no of data items are on increase. It can also be used to encode additional information by combining with other glyphs [13].

Shape Coding uses very small array pixels to visualize data. An array is used to represent one item of data.

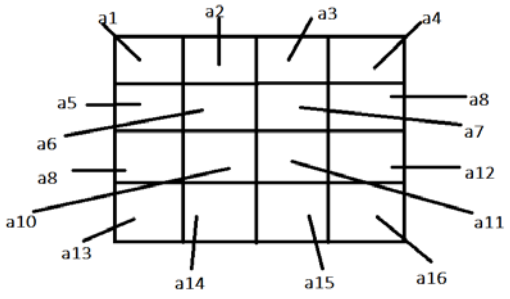

Fig. 4 Array for Shape Coding.(a1 to a16 are attributes)

4.9 Chart and Graph

These are most common, widely used and easily understandable ways to visualize information for the audience. Here are some of them:-

Line Graphs is also called line chart and shows relationship of a variable to other variable. It is used for trend tracking comparison of items within same period of time [27].

Bar Charts compare qualities of two or more groups. To show values bars, which can be either vertical or horizontal, are used. When there are large no of bars and same time bars are close together, it is not possible to detect differences between bars. That's why different colors are used for representing bars [27]. It works better when bars are having different range.

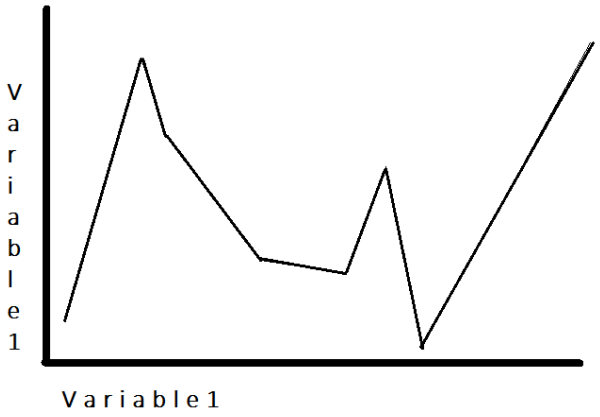

Fig. 5 Line Graph

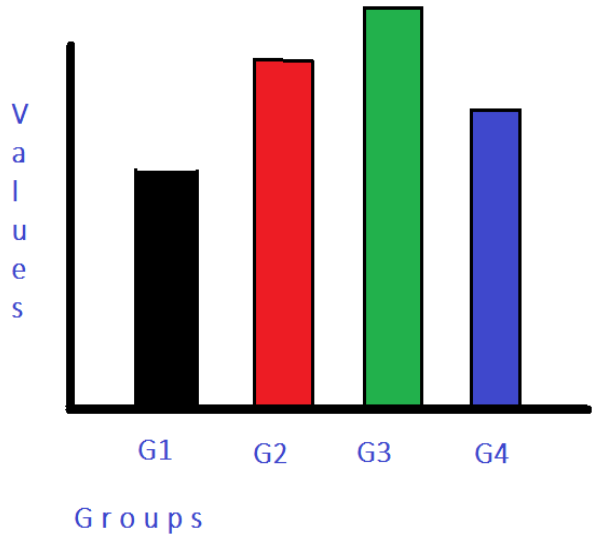

Fig. 6 Bar Chart (G1 to G4 are groups)

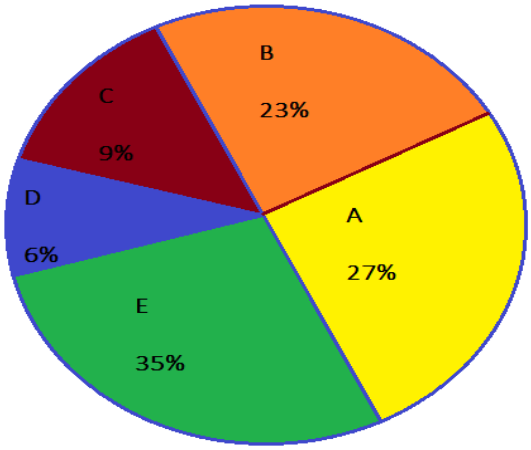

Fig. 7 Pie Chart

Pie Charts are subjects of discussion because their angles and areas cannot be easily interpreted by eyes. It is very useful when additional information (e.g percentage) is provided [27] and not ideal for developing dashboards for small screens.

5. Applications

Here are some of the applications of data visualization:-

1. **Business Decision Making Process: -** There are applications of visualization in the business decision making process. It enables the top level management to examine vast amount of data, find current markets trends, take the decision and make strategic changes if required. Common forms of visualization used in business decision making are basic charts, status indicators, scatter graphs, bubble charts, spark line charts, geographical maps, tree maps, Pareto charts etc [29].

2. **Other Areas Related to Decision Making Process: -**
   - **a) Uncertainty Visualization-** Uncertainty in the information is capable of influencing decision making. There are lots of techniques for uncertainty visualization.
   - **b) Risk Visualization-** Some problems also have risk in order to make decision to solve them. According to Lipkus & Hollands, (1999), users might wish to extract the following information [25] regarding risk: 1) Risk magnitude (i.e., how large or small the risk is);
     2) Relative risk (i.e., comparing the magnitude of two risks);

3) Cumulative risk (i.e., observing trends over time);
4) Uncertainty (e.g., estimating amount of uncertainty and variability or range of scores);
5) Interactions among risk factors.

Risk visualization uses static diagrams mostly.

c) Sensitivity analysis visualization- It uses graphs, charts, surface etc. There are very few techniques which can be applied to sensitivity analysis. Tornado diagram is a graphical approach which displays outcomes of local sensitivity analysis [25].

3. **Manipulate and Interact Directly with Data: -** Visualization enables users to directly interact and manipulate data unlike 1D table and chart which can only be viewed [38]. Real time visualization helps to figure out reasons for low performs of organization and can compare it with its rivals. And then most helpful changes can be made.

4. **Foster a New Business Language: -** Visualization tells all the things through data. Performance indicator does not tell about growing and shrinking category of business and the reasons behind it [38]. While visualization shows performance category wise and enable user to find reasons for it by further digging the data.

5. **Identify and Act on Emerging Trends Faster: -** Companies gather lot of data about their user by surveys, data mining and opinion analysis. Visualization is able to track [38] emerging trends and new opportunities for business related to those trends.

6. Xmdv Tool

Xmdv is one of the popular open source tools which are used for visualization process. It supports mainly 5 methods [39] listed below-

1. Scatterplots
2. Star Glyphs
3. Parallel Coordinates
4. Dimensional Stacking
5. Pixel-oriented Display

Applications areas of Xmdv tool includes remote sensing, financial, geochemical, census, and simulation data [39].

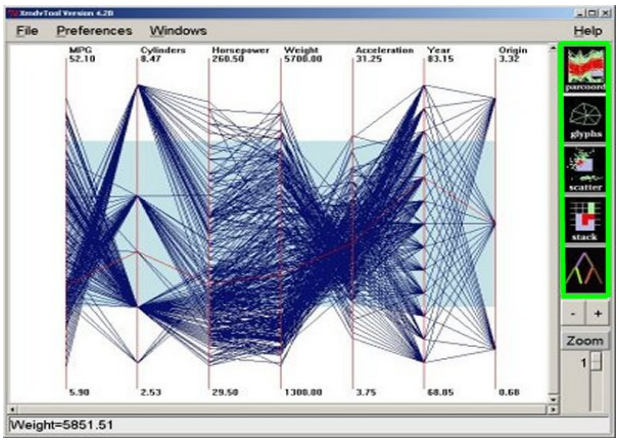

Fig. 8 Snapshot of Xmdv Tool.

References

[1] Denial Keim, Gennady Andrienko, Jean-Deniel Fekete, Carsten Gorg, Jorn Kohlhammer and Guy Melacon, "Visual Analytics: Definition, Process, and Challenges".
[2] Ilknur Icke, "Visual Analytics: A Multifaceted Overview".
[3] Ming C. Hao, Umeshwar Dayal and Daniel A.Keim, "Visual Analytics Techniques for Large Multi-Attribute Time Series Data".
[4] Zhao Kaidi, "Data Visualization".
[5] Jeffrey Heer, Michael Bostock and Vadim Ogievetsky, "A Tour through the Visualization Zoo".
[6] Martin Wattenberg, "Arc Diagrams: Visualizing Structure in Strings".
[7] Daniel A. Keim, Florian Mansmann, Daniela Oelke, and Hartmut Ziegler, "Visual Analytics: Combining Automated Discovery with Interactive Visualizations".
[8] Daniel A. Keim and Hans-Peter Kriegel, "Visualization Techniques for Mining Large Databases: A Comparison" IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, Dec. 1996.
[9] Doantam Phan, Ling Xiao, Ron Yeh, Pat Hanrahan and Terry Winograd, "Flow Map Layout".
[10] Joerg Meyer, Jim Thomas, Stephan Diehl, Brian Fisher, Daniel Keim, David Laidlaw, Silvia Miksch, Klaus Mueller, William Ribarsky, Bernhard Preim and Anders Ynnerman, "From Visualization to Visually Enabled Reasoning" Dagstuhl Seminar N° 07291 on "Scientific Visualization" - July 15-20, 2007.
[11] Pak Chung Wong, Chaomei Chen, Carsten Gorg, Ben Shneiderman, John Stasko and Jim Thomas, "Graph Analytics—Lessons Learned and Challenges Ahead".
[12] Melanie Tory and Torsten Moller, "Rethinking Visualization: A High-Level Taxonomy".
[13] Winnie Wing-Yi Chan, "A Survey on Multivariate Data Visualization".
[14] Ming C. Hao, Umeshwar Dayal, Daniel Keim, and Tobias Schreck, "A Visual Analysis of Multi-Attribute Data Using Pixel Matrix Displays".
[15] Adam Perer and Ben Shneiderman, "Integrating Statistics and Visualization: Case Studies of Gaining Clarity During Exploratory Data Analysis".

[16] Sandro Boccuzzo and Harald C. Gall, "Software Visualization with Audio Supported Cognitive Glyphs".
[17] Pak Chung Wong, Paul Whitney and Jim Thomas "Visualizing Association Rules for Text Mining".
[18] Ping Zhang and Andrew B. Whinston, "Business Information Visualization for Decision-Making Support -- A Research Strategy" Proceedings of the First Americas Conference on Information Systems, August 25-27, 1995, Pittsburgh, Pennsylvania.
[19] Danial A. Keim, Wolfgang Muller and Heidrun Schumann, "Visual Data Mining" Eurographics 2002.
[20] Svetlana Mansmann, Florian Mansmann, Marc H. Scholl and Daniel A. Keim, "Hierarchy-Driven Visual Exploration of Multidimensional Data Cubes"
[21] Michael D. Lee and Rachel E. Reilly, "An Empirical Evaluation of Chernoff Faces, Star Glyphs, and Spatial Visualizations for Binary Data" Australasian Symposium on Information Visualization, Adelaide, 2003. Conferences in Research and Practice in Information Technology, Vol 24.
[22] Tuan Pham, Rob Hess ,Crystal Ju, Eugene Zhang and Ronald Metoyer, "Visualization of Diversity in Large Multivariate Data Sets" IEEE Transactions on Visualization and Computer Graphics, Vol.16, No.6, November/December 2010.
[23] Martin S. Feather, Steven L. Cornford, James D. Kiper and Tim Menzies, "Experiences Using Visualization Techniques to Present Requirements, Risks to Them, and Options for Risk Mitigation".
[24] Mohammad Daradkeh, Clare Churcher and Alan McKinnon, "Supporting Informed Decision-Making Under Uncertainty and Risk through Interactive Visualization" Proceedings of the Fourteenth Australasian User Interface Conference (AUIC2013), Adelaide, Australia.
[25] Mohammad Kamel Younis Daradkeh, "Information Visualization to Support Informed Decision-Making Under Uncertainty and Risk".
[26] Stephen Few and Perceptual Edge, "Data Visualization Past, Present and Future".
[27] Justin Choy, Varsha Chawla and Lisa Whitman "Data Visualization Techniques from Basics to Big Data with SAS Visual Analytics", SAS Global Forum 2012 and SAS Global Forum 2011.
[28] Kristine Amari, "Techniques and Tools for Recovering and Analyzing Data from Volatile Memory", SANS Institute InfoSec Reading Room.
[29] Rebeckah Blewett, "The Importance of Data Visualization to Business Decision Making" June 12, 2011.
[30] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos and Samuel Kaski, "Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization".
[31] Ji Soo Yi, "Visualized Decision Making: Development and Application of Information Visualization Techniques to Improve Decision Quality of Nursing Home Choice".
[32] White Paper on "Big Data Visualization: Turning Big Data into Big Insights" Intel IT Center.
[33] Fernanda B. Viégas and Martin Wattenberg, "Artistic Data Visualization Beyond Visual Analytics".
[34] Wolfgang Müller and Heidrun Schumann, "Visualization Methods for Time Dependent Data- An Overview" Proceedings of the 2003 Winter Simulation Conference.
[35] Jens Lüssem, Stephan Schneider and Holger Studt, "Data Visualization Techniques" University of Applied Sciences Kiel, winter term 2013 / 14.
[36] Ming Hao, Umeshwar Dayal, Daniel Keim and Tobias Schreck, "Multi-Resolution Techniques for Visual Exploration of Large Time-Series Data" Eurogrpahics/ IEEE-VGTC Symposium on Visualization (2007), pp.1-8.
[37] Michael Balzer and Oliver Deussen, "Voronoi Treemaps".
[38] http://data-informed.com/top-5-business-benefits-using-data-visualization/
[39] http://davis.wpi.edu/xmdv/
[40] Fig. 8 Snapshot of Xmdv Tool. https://www.evl.uic.edu/cavern/multiperspective/Continuum/Training/xmdv1.jpg

Abhishek Kaushik is currently working in Siemens, Germany as a Master thesis student. He is in the final phase of completing his Masters degree in Information Technology from Kiel University of Applied Sciences. Before starting his Masters he received his Bachelor's of Technology in Computer Science Engineering from Kurukshetra University in 2012.

Sudhanshu Naithani has received his Bachelor's of Technology in Computer Science Engineering from Kurukshetra University in 2015. He is currently working as a research assistant under Assistant Professor Ravinder Madan at Manav Bharti University, Solan.