

Automate the chart validation with AI



Contents

Abbreviations	3
Introduction	4
Market trends	4
Problem statement	5
Solution	5
Case study	8
Benefits	10
Conclusion	11
References	11
Author info	12

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
CV	Computer Vision
DL	Deep Learning
CSV	Comma Separated Values
PDF	Portable Document Format
GB	Glucose Blood
DV	Data Validation
OCR	Optical Character Recognition
STLC	Software Testing lifecycle
HTML	Hyper Text Markup Language
ML	Machine Learning

Introduction

Graphs serve as an excellent interactive tool, allowing users across domains to compare, differentiate, categorise and visualise data easily. In an application or device's Software Testing Lifecycle (STLC), when the test output is represented as complex graphs, it can become tedious for testers to validate the graph contents manually. Various tools are available in the market to validate graphs against other graphs or to read values from external data sources for comparison. Multiple computations must be performed on external data in some scenarios to ensure accurate graph validations.

Graph validations encompass several aspects, including pixels, text, symbols, intersection points, color, shape and label. Scatter plot graphs, particularly, are complex charts that can feature multiple data points. This complexity increases when the dots and symbols used in the plots overlap or intersect, making manual validation challenging.

This whitepaper proposes a solution to automate complex graph validations using Computer Vision (CV) and deep learning techniques. The proposed solution enables the validation of complex scatter plot graph contents against various criteria and computations using baseline/reference files. It can precisely verify all the plots in the graph, regardless of the color, format or shape. This automated solution significantly reduces the manual efforts required for graph validations and enhances the testing cycle turnaround time.

Market trends

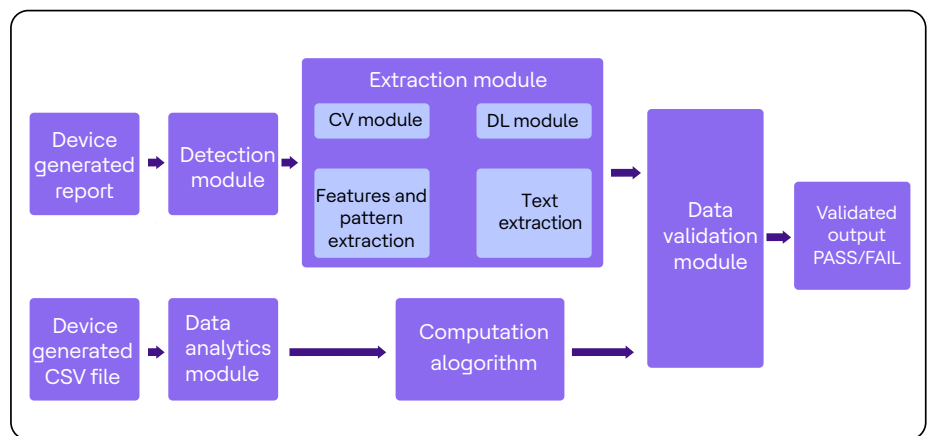
Companies are now turning to smart computer systems (AI) to check if their complex data connections, also known as graphs, are accurate. Harnessing the power of ML, these systems can automatically spot errors or unusual patterns in the data, accelerating the checking process and being more reliable compared to manual interpretation. Grand View Research states, "The global AI market size was valued at USD 196.63 billion in 2023 and is projected to expand at a compound annual growth rate (CAGR) of 37.3% from 2023 to 2030."

The world of digital data is growing fast and companies are turning to advanced tools, possibly including AI-driven graph validation, to handle complex information. This could be a key factor behind the significant growth forecasted in the global graph database market as businesses seek effective ways to manage and make sense of interconnected data. According to Markets and Markets, "The global graph database market size to grow from USD 2.9 billion in 2023 to USD 7.3 billion by 2028, at a CAGR of 20.2% during the forecast period."

Problem statement

- The customer device generates PDF reports accompanied by supporting data in CSV files (that contain values of various parameters).
- The primary challenge is validating the reports, which feature multiple graphs representing multiple parameters for a stipulated time duration.
- The tester has to perform computation manually on the supporting CSV files and needs to validate that each point plotted in the graph is precisely in line with the X and Y axis. It contains representations like (Min, max, average, ranges, circles, stars, circles with dots, etc.).

Solution



This whitepaper proposes a methodology for validating complex scatter plot graphs included in test reports along with their corresponding CSV file dump. The process begins when a test device generates test reports containing various graph images alongside supporting test information. The solution accommodates a multitude of test report formats, including PDF, Word, HTML, etc. Simultaneously, the test device generates a CSV data dump that contains extensive test information. The graphs in these reports may depict either single or multiple parameters based on the tests performed. Various representations in the charts, where some plots may indicate direct values for constants, ranges, maximums or minimums. This plotting can add visual complexity, especially when data points overlap during queries for timeframes ranging from 1 - 15 days, monthly, yearly or any stipulated period. Not all values in the CSV data dump will necessarily be available in the graph, as certain values may be excluded or included based on the parameter selection. This introduces additional complexity for testers, who must perform multiple validation on the graphs against the data contained in the CSV data file. The proposed whitepaper helps to read the appropriate values required for validating a particular plot in the graph without any manual intervention and with high precision using a combination of CV, DL and data analytics techniques.

The input to the solution will be the CSV file dump and report generated by the test device. The CSV dump file is sent to the data analytics module and the test report may be sent to the detection module. The graph part of the reports will be detected and extracted with their associated labels for further processing. The detected graph image is then passed to the extraction module to get the features and patterns through the CV module. The scatter plot graph image consists of numerous data points plotted, where each symbol will represent a unique parameter.

The scatter plot graph to be validated may have features in the background, like dashed grey lines connecting the plots, vertical and horizontal lines towards the X and Y axis, minimum and maximum range represented in horizontal yellow lines and minimum value to permissible range in the grey background (refer Figure 2), i.e., from 80 to 120 in Y-axis, the average values plotted in this range will be of blue circles '●' and beyond the range will be doughnut blue circles '◉'.

The star symbol '★' plotted will indicate the regular values observed during the time interval. The graph's X-axis will represent data collected for 15 days, 30 days, etc. or any stipulated duration with 24-hour intervals.

The Y-axis will represent the GB reading that includes a scale range from 20 to 280 and the GB reading below 20 gets plotted on the scale 20 using red rhombus '◆'; likewise yellow rhombus '◇' on the scale 280 for values above 280. The average values and the number of readings during the time interval concerning the X-axis will be available on top of the graph.

The objective is to validate all the points plotted, such as a star, rhombus, blue circle and doughnut blue circle, concerning X and Y-axis values. The extraction process of all the symbols used for parameter representation precisely alongside their coordinates is challenging. The table presented above, the graph that contains average values and the number of outputs between the time intervals will also be validated.

The number of star outputs present for each time interval count will be proportional to the value displayed in the 'No. of Outputs' row in the table. Likewise, the average value of the star plot value will be present in the table for each X-axis interval. Both the number of outputs and average values will be validated against the values from the CSV file. This type of verification on a graph with high precision will make a manual tester job a tedious and time-consuming task. The reference values between the scale pointers for intermediate values are not available for the X and Y-axis interval for reference, which makes it more challenging and additional computations need to be performed by the tester. Various pre-processing techniques are applied to the graph to differentiate the background, remove noise and enhance the image to get the plots precisely detected. Given that the graph represents various parameters, separate shapes are employed to symbolize each parameter. The text contents in the identified graph image are processed using OCR techniques in the DL module for reference and validation. The extracted features and text get passed to the validation module.

The CSV data file dump gets passed to the analytics module that contains data values such as time, GB reading, date, device information, time range to be included and excluded, etc. At the same time, in the graph, the representation will be for 24-hour intervals for a date range of 1-15 days. So, there is a possibility that each day, the values will be replicated in the same time frame. For example, days 1, 2 and 5 will have a value on the X-axis at 2:24 and GB reading as 60, so all the plots will be overlapping and representing as a single plot. The GB readings below the minimum value of 20 will be plotted on the Y-axis on the same scale/line. The data analytics engine identifies those days 1, 2 and 5 have three of the same values as entries in the CSV file dump. Still, it must be considered as a single entity concerning validation. The data is to be analysed and filtered to get average GB values and the number of outputs and validated with the values displayed on top of the graph image concerning time intervals of the X-axis. Various computations need to be performed on the CSV file to validate each plot represented in the graph. This scenario includes challenges in validation using CV techniques when points are intersecting, partially overlapping, the same color, etc.

The computed values for each parameter to be validated get passed to the DV module to verify the plots in the graph with data. This module will extract the features and text from the graph, along with its coordinate details on the X and Y-axis. The computed values from the CSV file are available in the DV module for verification. Based on parameter values from the DV module, the coordinates will be identified and marked in the graph with respect to the X and Y-axis. The marking done by the validation engine for internal use for each point should be present inside the symbol (rhombus, circle, doughnut circle, star, etc.) used for representation. Based on this, the solution will be verified internally and declared as a 'Verification Pass'. If the market value is outside the symbol representation, it will be considered a verification fail. That particular symbol will be highlighted with a red box. The additional validation includes the number of plots present between each time interval to be calculated and verified with the value in the table on top of the graph.

The average value for GB reading for time intervals will be calculated and examined using the table above the graph. The minimum range value of 20 will be extracted from the Y-axis and verified with all plots below the range or plotted in that coordinate, likewise for the maximum value. The values below the minimum range of 20 and values above the maximum range of 280 are plotted using the rhombus symbol by the test device to get validated. The process of verifying the rhombus symbol is done by checking if the markings made using CSV values fall within the respective region of the symbol. Average values of each time interval corresponding to the X-axis are plotted and validated against the coordinates of the blue and blue doughnut circles. The average values and number of GB readings plotted per interval are verified with the text extracted from the table above the graph image. On performing all the required validations in the graph with the CSV file, the solution will highlight the average values in a red box for verification fail cases. The user can mouse hover over a specified time interval range and a pop-up window will be displayed with the CSV and plotted values. It helps the user to know the difference between the actual and the plotted. The color of the pop-up changes according to the result, such as green if validation passes and red if validation fails.

Case study

Below is a sample of the data from the CSV file dump (Table 1). It can be in any format, such as CSV, Excel, etc. It contains contents like index number, time, date and output reading. These are the values to be verified with the graph image. The time is represented as the X-axis values and the GB reading is represented as the Y-axis values. All values will be plotted in the same graph for 15 or 30 days.

Index	Date	Time	GB Reading	D-name	D-stat
0	1/10/2023	12:46 AM	80	PIAB1	1
1	3/10/2023	1:15 AM	295	PIAB1	1
2	6/10/2023	1:15 AM	299	PIAB1	1
3	7/10/2023	4:10 AM	345	PIAB1	1
4	8/10/2023	6:30 AM	234	PIAB1	1
5	8/10/2023	8:47 AM	126	PIAB1	1
6	8/10/2023	12:05 PM	184	PIAB1	0
7	10/10/2023	4:35 PM	95	PIAB1	0
8	11/10/2023	7:22 PM	278	PIAB1	1
9	14/10/2023	9:55 PM	200	PIAB1	1
10	15/10/2023	6:45 PM	328	PIAB1	1
11	15/10/2023	3:33 PM	222	PIAB1	1

Table 1: Table from CSV file data

The values from this table will be validated with the respective parameters from the graph image. CV techniques plot these values in the graph image for internal reference.

The graph shown below (Fig 2) is a scatter plot graph; the X-axis consists of the time scale of 24 hours and the Y-axis consists of the GB reading. GB reading has a minimum range of 20 and a maximum of 280. In the Y-axis reading, if any plot goes below this minimum range, it is plotted in red rhombus on the origin line and if the plot goes above the maximum range, it is plotted in yellow rhombus. The plot symbol could be any symbol such as a circle, star, square, etc.; the graph used in this whitepaper has star symbols as plots. The stars are plotted from the time and GB reading and the doughnut circles are the average value for every time interval in the X-axis; the circle will be fully shaded when the average value falls under a range of 80-120. The red horizontal line at 80 GB denotes the minimum range. The graph image also contains average values and the number of outputs for each X-axis interval in a table on top of it.

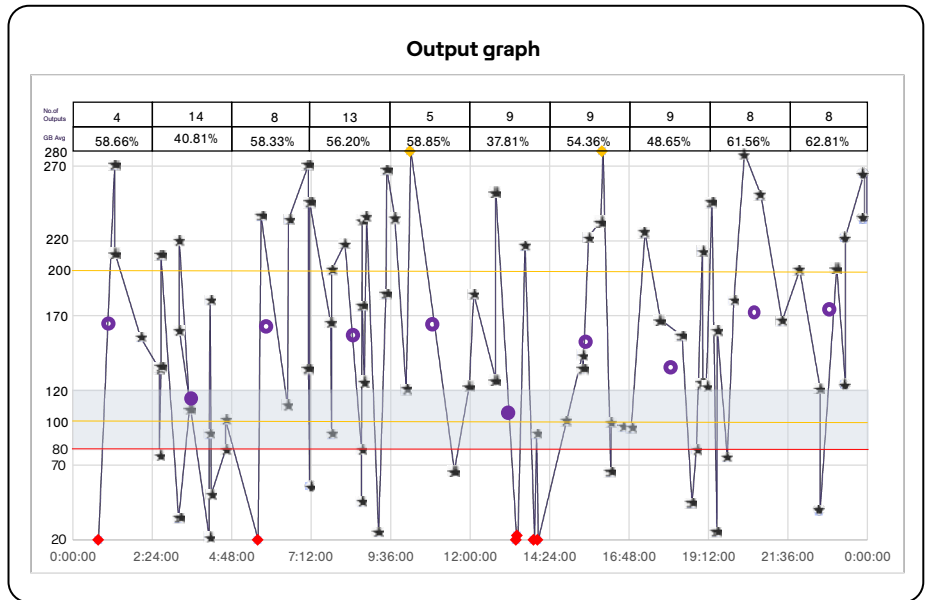


Fig 2: Output graph image

The validation process begins by verifying whether the plots marked from the CSV file align with the corresponding elements within the graph image.

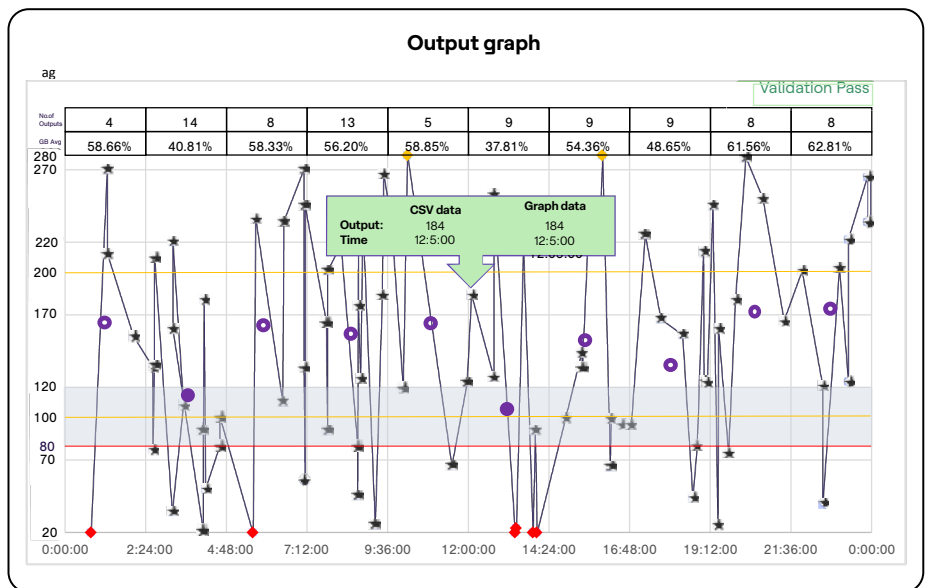


Fig 3: Validated output

The above graph (Fig 3) shows the result after validating values from the graph and CSV data file. The graph below (Fig 4) shows the negative scenario where both the data don't match, i.e., the CSV value does not lie in the region of its respective star symbol. The plots that don't satisfy the verification process, i.e., if the value from the CSV file does not come into the star plot's area, will be highlighted with a red box for user convenience.

The average value will also be calculated according to the plots and will be highlighted in red if both content from the graph value and CSV file do not match. The validation process begins by verifying whether the plots marked from the CSV file align with the corresponding elements within the graph image. For all the matching plots, the annotation window will pop up while the mouse hovers over it with the time and output values in a green color pop-up box. When the plots do not match, the pop-up box will be red with the time and output values.

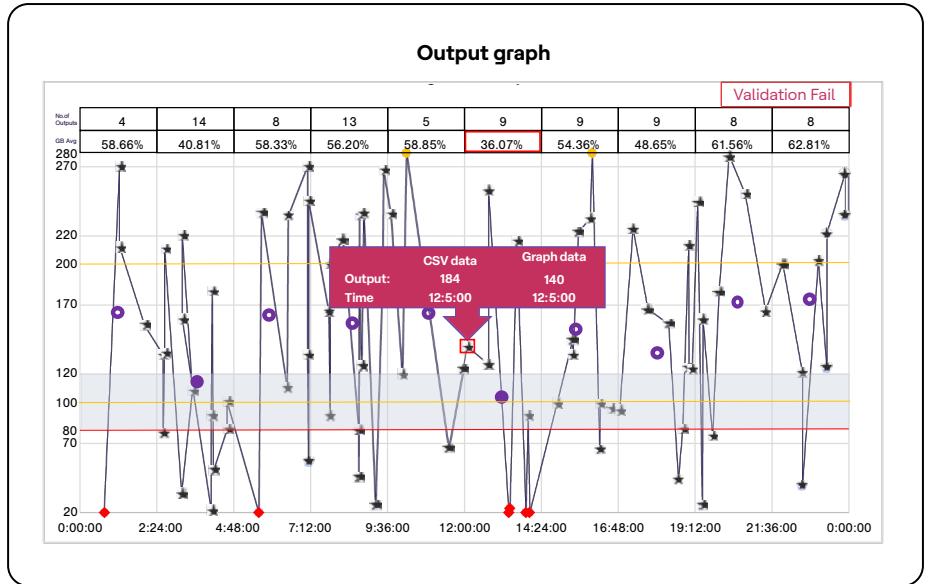


Fig 4: Negative scenario output

Benefits

The proposed solution offers several benefits, including:

- **Reduced time consumption:** The solution automates the graph validation process, significantly decreasing the time and efforts required for manual testing.
- **Improved accuracy:** The solution accurately validates the graph's contents by employing CV and DL techniques, minimizing the risks of errors and defects.
- **Enhanced scalability:** The solution is capable of handling complex graphs with multiple plots and various color formats, making it more scalable than manual testing.
- **Flexibility:** The solution can be customized to accommodate various graphs and validation methodologies, providing greater flexibility compared to manual testing.

Conclusion

This proposed solution addresses the challenges associated with manually validating graphs by automating the validation process by comparing graph contents against a CSV file dump and determining pass or fail outcomes. Both the graph and CSV data file serve as input for this validation. The validation process of complex graphs like dot plot graphs can be automated, which helps to reduce the time taken for the manual validation process and with more accuracy. This solution uses techniques from AI, CV and DL. With this solution, the validation process of the dot plot graph can be made simpler and easier.

References

- [What is an OCR ???. A basic theoretical overview of the... | by Susmith Reddy | Towards Data Science](#)
- [Computer Vision Tutorial \(geeksforgeeks.org\)](#)
- [Artificial Intelligence Market Size And Share Report, 2030 \(grand-viewresearch.com\)](#)
- [Graph Database Market Size, Industry Share, Emerging Trends & Opportunities | MarketsandMarkets™](#)

Author info



Narender S

Narender S has been in the software engineering industry for over 15 years. Additionally, he has several years of experience in product engineering and sustenance engineering across domains. He has a strong understanding of the software automation process and has a proven track record of delivering results for enterprise clients.



Srihari V

Srihari V has been in Telecom and networking for the past 16 years. He has managed various testing teams and has been creating next-gen solutions as value-adds for leading OEM clients. He is currently part of the Solutions team at HCLTech and generates AI-based solutions to support business needs.



Sri Gayathri Paaraa P

Gayathri holds a degree in Computer Science and Engineering. She possesses two years of experience in working with AI, ML, image processing, deep learning and computer vision techniques.

HCLTech | Supercharging Progress™

HCLTech is a global technology company, home to 222,000+ people across 60 countries, delivering industry-leading capabilities centered around Digital, Engineering and Cloud powered by a broad portfolio of technology services and software. The company generated consolidated revenues of \$12.3 billion over the 12 months ended December 2022. To learn how we can supercharge progress for you, visit hcltech.com.

hcltech.com

